A Comprehensive Analysis of Human CpG Island Methylation

Robert S. Illingworth

Thesis presented for the degree of
Doctor of Philosophy

The University of Edinburgh

2008

Dedicated to G.T.L. (Jock) Scott

Acknowledgements

I met Adrian Bird at the conclusion of my undergraduate studies, and he is largely to blame for my current situation! Thanks must primarily be given to you Adrian for giving me the opportunity to work and study in your lab, and for your enthusiasm which has spurred me on throughout my PhD! I would like to thank Christine Struthers for excellent and tireless administrative help. From the lab I would particularly like to thank Heather Owen for the provision of constructive criticism on this thesis, Rodoniki Athanasiadou for help with bioinformatic analysis and Aileen Greig who keeps the lab running smoothly. Moreover I would like to thank Dina DeSousa and Elaine Evans for excellent assistance with Bisulfite sequencing! I thank all remaining lab members for a combination of insightful scientific discussion and a lot of laughs along the way.

Within the WTCCB, I owe greatest thanks to Alastair Kerr who has provided extensive Bioinformatic support and advice throughout my project. I would also like to thank Jose de Las Heras for considerable help with microarray analysis and to Ahmed Raza who provided me with a continuous source of human blood!

There have been many external contributors to my work, none more critical to my success than members of staff at the Wellcome Trust Sanger Institute. Cordelia Langford, Peter Ellis and colleagues generated the sequence data and microarrays which were central to this project.

Finally, thanks to those people outwith academia that have been instrumental in getting me to this point. To my parents, Sue and David I thank you for your unwavering support during my many years as an unemployed student! Irene and Colin, who have provided both a roof over my head and many hearty meals over the past few years. To my friends Duncan, Karen, Sharon, Scott, Aaron, Gill, Chis and the rest who have kept me sane (relatively!) throughout the last few years. Lastly, and by no means least, I thank my fiancée, Louise Brady, for support and tolerance, particularly during the past few months - you are a 'catch'!

I would also like to thank the MRC for PhD funding, without which I would probably be living in a box in the Meadows!

Abstract

Global methylation of the human genome has led to the depletion of CpG dinucleotides. CpG islands (CGIs) represent a conspicuous exception to this rule as they generally lack cytosine methylation and are characterised by clustered CpG dinucleotides. These sequences are non-randomly distributed throughout the genome, colocalising with approximately 60% of gene promoters. CGIs generally associate with transcription start sites (TSSs), a small proportion of which are normally methylated and correlate with gene inactivity. This phenomenon is best characterised in the context of X-inactivation, where the majority of CGIs are specifically methylated on one of the two X chromosomes. Aberrant CGI methylation is also a common feature of neoplastic cells and has been suggested to play a role in tumour development and propagation.

We devised a preparative chromatographic tool for the enrichment of CGI sequences based on the empirical criterion of clustered nonmethylated CpGs. Characterisation of a novel somatic CGI set derived from human blood DNA identified a population of islands which were missed by commonly applied bioinformatic criterion. The identified CGIs were enriched in gene rich portions of the genome, and preferentially associated with chromosome ends proximal to the telomeres. Despite association with gene promoters, approximately fifty percent of CGIs were found to be either intra- or intergenic. An arrayed set, comprising ~17,000 unique sequences, allowed the investigation of the methylated CGI component of human blood, brain, muscle and spleen. This identified 6-8% of normally methylated CGIs across the 22 human autosomes. These hypermethylated islands were enriched in regions distal to annotated promoters and preferentially associated with developmental gene loci. The reagents and observations discussed in this thesis allow detailed analysis of the abundance and distribution of methylated CGIs in 'normal' human somatic cells.

Table of Contents

Figures

Tables

DVD contents

File Name	Title	File format
Dataset 1	Unique CpG Islands	.txt
Dataset 2	Filtered CGI set for Microarray analysis	.txt
Dataset 3	Classification of Methylated CGIs	.xls

Abbreviations

5AzaC	5-Azacytidine
5MeC	5-Methylcytosine
Ac	Acetyl-
C5/6	Carbon 5 / 6
CAP	CXXC Affinity Purification
CGBP	CpG Binding Protein
CGI	CpG Island
ChIP	Chromatin Immunoprecipitation
CIMP	CpG Methylator Phenotype
CNBr	Cyanogen Bromide
DMR	Differentially Methylated Region
DNMT	DNA Methyltransferase
dpc	Days post coitus
dsDNA	Double stranded DNA
E	Embryonic Day
ES	Embryonic Stem
GO	Gene Ontology
H	Histone
HDAC	Histone Deacetylase
HMTase	Histone Methyltransferase
HP1	Heterochromatin Protein 1
HTFs	HpaII Tiny Fragments
ICF	Immunodeficiency, centromeric heterochromatin and facial abnormalities
IPTG	Isopropyl-β-D-thiogalactoside
K	Lysine
lmPCR	Linker Mediated PCR
MAGE	Melanoma Antigen Encoding Genes
MAO	Monoamine Oxidase
MAP	MBD Affinity Purification
MBD	Methyl-CpG-binding domain
MBP	Methyl-binding Protein
me	Methyl-

MeCP2	Methyl-CpG-binding protein 2
meCpG	Methyl-CpG
MEF	Mouse Embryonic Fibroblast
ncRNA	non-coding RNA
NLS	Nuclear Localisation Centre
NPC	Neuronal Progenitor Cell
ORF	Open Reading Frame
Pc	Polycomb
PcG	Polycomb Group
PCNA	Proceeding Nuclear Antigen
PGCs	Primordial Germ Cells
ph	Phosphoryl-
PRC1	Polycomb Repressive Complex 1
PRC2	Polycomb Repressive Complex 2
RdDM	RNA Directed DNA Methylation
rDNA	Ribosomal DNA
RLGS	Restriction Landmark Genomic Scanning
RNAi	RNA interference
rRNA	Ribosomal RNA
SAM	S-adenosyl-L-Methionine
siRNA	short interfering RNA
SNP	Single Nucleotide Polymorphism
ssDNA	Single Stranded DNA
TFs	Transcription Factors
TGS	Transcriptional Gene Silencing
TRD	Transcriptional Repressor Domain
TrxG	Trithorax Group
TSA	Trichostatin
TSS	Transcription Start Site
Xa	Active X Chromosome
Xi	Inactive X Chromosome
Xic	X-inactivation Centre

Chapter 1: Introduction

Approximately 40 rounds of cell division are required to generate the $10x10^{13}$ cells comprising the human body. These can be subdivided into more than 700 phenotypically distinct cell types, the function and identity of which are dictated by the specific expression pattern of approximately 22,000 protein coding genes. These patterns are maintained via a highly orchestrated transcriptional regulatory system following cell type specification. The mechanisms governing the identity of these cells are highly complex, requiring cell specific signals to translate the same genetic material into the protein composition required for specific cellular functions. Despite the fact that this system is regulated by sequence specific transcription factors, these alone cannot account for the functional diversity of these genetically identical human cells. Accordingly, non-genetic information, in the form of epigenetic modifications, is also required for correct gene regulation.

1.1 Epigenetics

The prefix Epi- describes an event which occurs 'upon' and therefore Epigenetic literally translates into 'upon the gene'. Epigenetics, in its simplest form, refers to any modulation of cellular phenotype which is not directly encoded in the DNA sequence. A recent definition states: "the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states" (Bird, 2007). This includes various contemporary features of cell biology such as the Polycomb and Trithorax system, histone modifications and DNA methylation.

The initial sequencing of the human genome seven years ago, heralded a new era for global sequence based analysis (Lander et al., 2001). In particular, this has allowed the generation of DNA microarrays which have been applied to the dissection of epigenetic features on a genome wide scale. Moreover, programmes such as the HEP (Human Epigenome Project), HEROIC (High-throughput Epigenetic Regulatory Organisation In Chromatin) and ENCODE (ENCyclopedia Of DNA Elements) have provided extensive information regarding the global distribution of human epigenetic components (Birney et al., 2007; Eckhardt et al., 2004; Eckhardt et al., 2006; ENCODE Project Consortium, 2004).

Epigenetic mechanisms serve to increase the coding potential of the DNA sequence by providing information regarding an activity state, either past or present. Transcription in small bacterial genomes is regulated principally by sequence specific transcription factors, which bind their cognate sites and elicit an effect. Eukaryotic genomes are many orders of

magnitudes bigger than their bacterial counterparts and may have evolved epigenetic mechanisms to reduce the effective complexity of the genome. Condensed chromatin, as dictated by epigenetic factors, restricts access to the majority of the non coding portion of the genome, and consequently facilitates association of transcriptional machinery to their cognate sites. This concept has led to the idea that chromatin serves as a modulator of transcriptional permissivity. However, genetic studies have determines that various core components of epigenetic mechanisms are essential in mammals (Li et al., 1992; Okano et al., 1999). This suggests that epigenetic systems may not merely 'fine tune' the genome but rather play a fundamental role in its architecture and gene regulation. Consequently, chromatin modifications, polycomb/Trithorax proteins and DNA methylation have all been investigated in a bid to elucidate their function in genomic regulation. The following section provides a brief overview of chromatin structure to provide a context for the discussion of epigenetic systems discussed in this thesis.

1.1.1 Chromatin Modulation

Eukaryotes package their genetic material into a nucleoprotein structure referred to as chromatin. Chromatin provides a structure which can be modulated to elicit compaction (up to 10,000 fold) or accessibility of the DNA as is required for specific biological processes (Fig. 1.1-1). The nucleosome represents the core component of this structure which is composed of approximately 150bp of DNA wrapped around an octamer of four conserved histone proteins (H3, H4, H2A and H2B; (Luger et al., 1997)). The fundamental role for this structure is to spatially and temporally modulate DNA accessibility to facilitate cellular processes (Kouzarides, 2007; Li et al., 2007a). What mechanisms mediate the physical properties of chromatin to afford such dynamic functionality?

Covalent Chromatin Modifications

Histones are largely globular proteins with the exception of their highly unstructured amino terminal tails which protrude past the DNA (Luger et al., 1997). These tails serve as a scaffold on which numerous posttranslational modifications can be bestowed. These modifications include; Methylation (me), Phosphorylation (ph), Acetylation (ac); ADP ribosylation (ar), Ubiquitination (ub) and Sumoylation (su). In excess of 100 modifications have been identified, owing to their ability to occur on multiple amino acids and, in the case of methylation, in -mono, -di, and -tri methylated forms (Bernstein et al., 2007; Kouzarides, 2007).

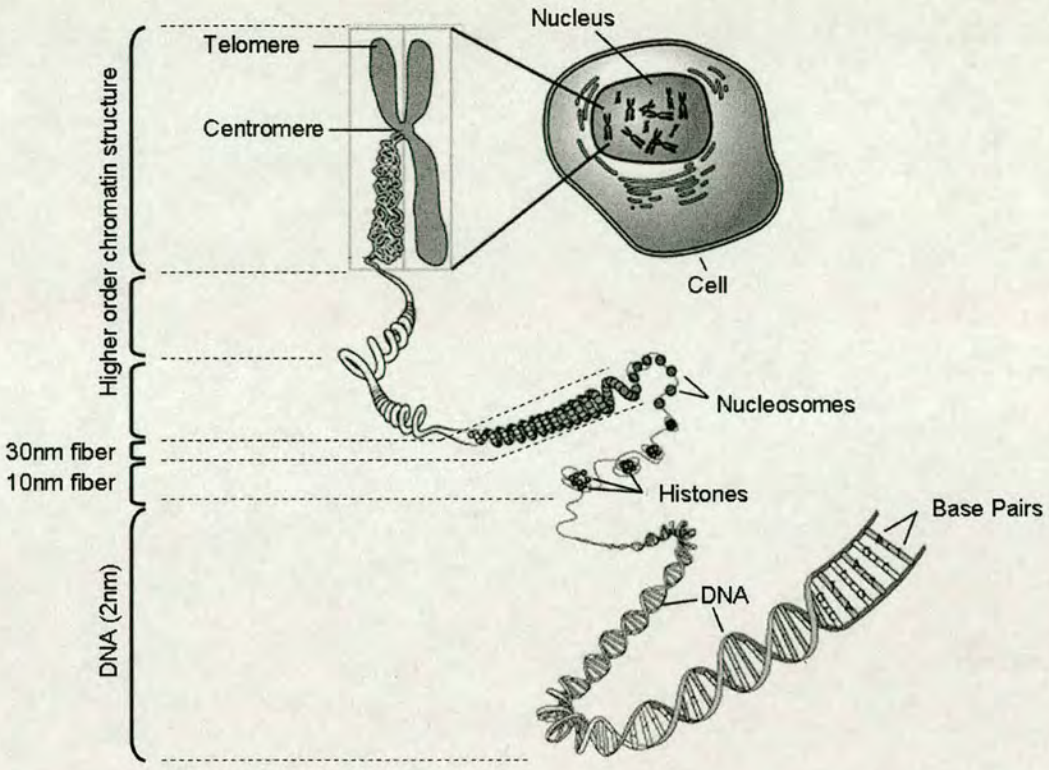


Figure 1.1-1. Compaction of nuclear DNA
Eukaryotic DNA is assembled into chromatin, by superhelical wrapping around histone octamers to comprise nucleosomes. The 10nm fiber is further compacted into higher order structures which comprise s chromosomal heterochromatin (Figure adapted from http://employees.csbsju .edu/hjakubowski/classes/ch331/dna/chromosome .gif)

These modifications can largely be partitioned into two distinct categories representing open transcriptionally permissive and restrictive transcriptionally inert chromatin conformations. With respect to transcription, H3K4me2, H3K4me3, H3K36me3, H3K27me1, H3K9me1, H3/H4ac, H4K20me1, H2BK5me1 and H2Bub1 have all been associated with transcriptional activity in mammalian cells (Barski et al., 2007; Bernstein et al., 2005; Kouzarides, 2007; Li et al., 2007a; Roh et al., 2005). Alternatively, H3K9me2, H3K9me3, H3K27me3 and hypoacetylation of H3/H4 have been associated with transcriptional silencing (Barski et al., 2007; Kouzarides, 2007). This is not an exhaustive list and additional modifications appear to be involved in both activation and repression in a context specific manner (Barski et al., 2007; Bernstein et al., 2005; Kouzarides, 2007; Li et al., 2007a; Roh et al., 2005). Acetylation effects transcription by directly altering the physical characteristics of chromatin. The acetyl groups neutralise the positive charge of the histone tails, destabilizing inter-nucleosome interactions and facilitating chromatin decondensation (Luger et al., 1997). Alternatively, histone modifications can provide binding sites which recruit protein factors

and facilitate or antagonize transcription. Methylation of H3K9 and H3K27 recruits heterochromatin protein1 (HP1) and Polycomb (Pc) respectively which maintain transcriptional silencing through formation of a repressive chromatin environment (Bannister et al., 2001; Cao and Zhang, 2004; Lachner et al., 2001). Alternatively, H3K4 methylation directly associates with the basal transcriptional machinery in mammals whilst abrogating DNMT (DNA Methyltransferase) binding to facilitate transcriptional activity (Ooi et al., 2007; Vermeulen et al., 2007). These modifications can mediate cross-talk between different epigenetic systems. This is highlighted by the requirement for H3K9me3 in *N. crassa* which is essential for establishing DNA methylation patterns (Tamaru et al., 2003).

The existence of a histone code has been proposed, which implies that a combination of posttranslational modifications will dictate and therefore predict a transcriptional state (Margueron et al., 2005; Nightingale et al., 2006). In light of these observations this proposition is logical; however it provides a somewhat 'black and white' view of chromatin regulation. Modifications which are often associated with transcriptional activity or inertia do not always concord with the observed transcriptional state (Bernstein et al., 2006). Moreover, this notion does not directly account for transcriptional factors which fundamentally govern gene expression levels. This is not to say that epigenetic modifications do not underpin transcriptional regulation. Indeed if this was not the case, treatments with pharmacological agents such as 5-Azacytidine (5AzaC) and TSA (Trichostatin A) would have little effect on global expression patterns, which is not consistent with experimental observations (Hansen and Gartler, 1990; Mohandas et al., 1981; Yoshida et al., 1995). The evidence suggests that histone modifications facilitate or abrogate access to functionally important DNA sequences. These would serve to reduce the complexity of large genomes, and mediate the correct localisation of transcription factors to their cognate sites (Bird, 2002).

Chromatin remodeling and Histone Replacement

Interaction between sequence specific protein factors and DNA can be precluded by nucleosomes (Saha et al., 2006; Varga-Weisz, 2001). In the 'ground state', nucleosome positioning is largely dictated by DNA sequence composition, which provides thermodynamically favorable localisation sites (Ioshikhes et al., 2006; Satchwell et al., 1986). However, nucleosome translocation along DNA can allow specific DNA sequence to become accessible or occluded for protein factor interactions which can effect gene expression (Muchardt and Yaniv, 1999; Saha et al., 2006; Sudarsanam et al., 2000). This

dynamic mechanism is provided by ATP-dependent nucleosomal remodeling, intrinsic to several families of multisubunit protein complexes. Mammalian co-repressor complexes such as NuRD, contain chromatin remodeling components required for transcriptional repression (Le Guezennec et al., 2006; Zhang et al., 1999).

Chromatin remodeling has also been implicated in replication independent histone replacement, whereby canonical histones are substituted for variants in response to cellular triggers including transcription and DNA damage (Mizuguchi et al., 2004). H3.3 is a variant H3 isoform which differs from H3.1 (the major form) by 4 amino acids. Unlike H3.1, this variant can be deposited onto active chromatin independently of DNA replication (Ahmad and Henikoff, 2002). This has been suggested to facilitate transcription via the erasure of inactivating chromatin modifications such H3K9 methylation (Ahmad and Henikoff, 2002; Janicki et al., 2004). A second H3 isoform, CENPA, is a component of centromere specific nucleosomes. CENPA is required for the establishment of centromeric heterochromatin, and provides a scaffold for the recruitment of kinetochore proteins involved in spindle formation and chromosome segregation during mitosis (Smith, 2002). The H2A variants H2AZ and H2AX facilitate nucleosome expulsion and DNA repair respectively. H2AZ, is localised to the promoter proximal regions of active genes, and has been proposed to have a role in destabilizing chromatin to facilitate transcription ((Schones et al., 2008) and references therein). H2AX localizes throughout the genome and becomes rapidly phosphorylated in response to double strand breaks (Redon et al., 2002). Possibly, one of the most striking examples of histone replacement occurs during the formation of the male germ line. Spermatocyte heterochromatin is restructured by replacing histones with protamines. Protamines are unrelated to histones and result in the formation of a distinct, highly compacted nucleoprotein structure which is highly condensed with respect to somatic chromatin (Wouters-Tyrou et al., 1998). Both chromatin remodeling and nucleosome replacement provide mechanisms by which chromatin structure can be modulated in response to cellular signals.

1.1.2 Polycomb and Trithorax proteins

Polycomb (PcG) and Trithorax (TrxG) proteins represent two classes of antagonistic factors which are highly conserved in metazoans. They were first characterised based on mutational phenotypes which resembled those of homeotic genes in *Drosophila* (Kennison, 1995). They function to regulate and memorise transcriptional activity of developmental genes, and maintain the expression state even when the initial cues are diminished. They exert their

influence on transcription by modulating the activity of specific chromatin loci through histone modification and nucleosome remodeling. Such mechanisms are required to maintain cellular identity through maintenance of appropriate gene activity in differentiated cells (Cao and Zhang, 2004; Schuettengruber et al., 2007).

PcG proteins are negative regulators of developmental gene expression, and are broadly categorised into two repressive multisubunit complexes. The PRC1 (Polycomb repressive complex 1) contains the archetypal Polycomb protein (HPC1-3)[i], HPH1-3 (Polyhomeoitic), BMI1 (BMI1 polycomb ring finger oncogene) and RING1A/B. The RING proteins can modify histone H2A by the addition of ubiquitin on K119 (de Napoles et al., 2004). The second complex is PRC2 (Polycomb repressive complex 2) which catalyses the formation of H3K27me2 and me3 (Cao and Zhang, 2004). This activity is provided by EZH2 (Enhancer of Zeste homologue 2) which in combination with EED (Embryo Ectoderm Development), SUZ12 (Suppressor of Zeste 12) and RbAp48 comprise the core of the PRC2 complex (Cao and Zhang, 2004; Schuettengruber et al., 2007).

PcG proteins are required to mediate the correct expression of lineage specific genes during cellular differentiation (Cao and Zhang, 2004; Chamberlain et al., 2008). Consistent with this, genes which associate with SUZ12 in human Embryonic Stem (ES) cells were preferentially activated during artificial differentiation (Lee et al., 2006). Tissue specific genes not required in a particular cell lineage maintained SUZ12 association at the promoter consistent with transcriptional inactivity (Lee et al., 2006). PRC1 can be recruited to chromatin bearing H3K27 methylation via interaction with the chromodomain of the polycomb protein. This interaction provides a mechanism whereby the activity of PRC2 can facilitate PRC1 association with chromatin (Cao and Zhang, 2004; Schuettengruber et al., 2007).

The ncRNA (non-coding RNA) *HOTAIR*, was recently shown to recruit PRC2 to the *HOXD* locus and was required for H3K27 methylation (Rinn et al., 2007). During mammalian X inactivation, both PRC1 and PRC2 are recruited to the inactive X chromosome (de Napoles et al., 2004; Plath et al., 2003). However PRC1 can associate independently of PRC2 via an interaction with the ncRNA; *Xist* (The role of PcG in X inactivation will be discussed later; (Schoeftner et al., 2006). These data suggest that PRC1 and PRC2 recruitment can be facilitated by interaction with RNA molecules.

[i] All PcG and TrxG subunits are named as per the human nomenclature.

TrxG proteins are involved in memorising and propagating transcriptional activity. In humans, the SWI/SNF and NURF TrxG complexes provide remodeling activities whereas the MLL1-3 complex contains a histone methyl transferase (HMTase) (Schuettengruber et al., 2007). The *in vitro* chromatin remodeling activity of the SWI/SNF complex was shown to be abolished by purified PRC1 from *Drosophila* extracts (Shao et al., 1999). MLL contains a SET (Su(var)3-9, Enhancer-of-zeste, Trithorax) domain which specifically methylates H3K4 (Dou et al., 2005; Nakamura et al., 2002). Interestingly, both the MLL and CFP1 (CGBP) components of the MLL1-3 complex possess CXXC motifs which have been shown to bind DNA containing non-methylated CpG sites (Birke et al., 2002; Voo et al., 2000). A recent study, investigating histone methylation in murine ES cells, indicated that the vast majority of CpG Island (CGI) promoters were trimethylated at H3K4 sites (Mikkelsen et al., 2007). Moreover TFIID, a multisubunit component of the mammalian basal transcription machinery, has been shown to interact with trimethylated H3K4 (Vermeulen et al., 2007). This suggests a possible link between the underlying DNA methylation status and activation of developmental genes.

Recent, analysis of conserved regions of the mouse genome led to the identification of 'bivalent chromatin domains', comprising large domains of H3K27me3, over-laid with more discrete regions of H3K4me3 (Bernstein et al., 2006; Mikkelsen et al., 2007). Consistent with the functional role of the PcG and TrxG complexes which distribute the modifications; these regions were found to associate with cell type specific genes. It has been proposed that the combination of these opposing marks, maintains these genes in a poised state to be activated or repressed during linage specification. These domains generally resolve into either the K4 or K27 methylated states which correspond with gene activity and repression respectively (Fig. 1.1-1; (Bernstein et al., 2006; Mikkelsen et al., 2007)).

Recently, there has been extensive characterisation of chromatin modifying enzymes which can specifically demethylate lysine residues on histone tails (reviewed in (Klose and Zhang, 2007)). Such activities could conceivably facilitate the resolution of these domains during cellular differentiation. The TrxG protein UTX (Ultratrithorax) was characterised as an H3K27 di- and trimethyl specific demethylase (Agger et al., 2007). Moreover, proteomic analysis determined that UTX could physically associate with the MLL histone methyltransferase complex in HEK293 nuclear extracts (Lee et al., 2007a). This interaction suggests a link between H3K4 methylation and H3K27 demethylation during development

(Agger et al., 2007; Lee et al., 2007a). Conversely, the Drosophila Lid (Little Imaginal Disc) protein can catalyse the removal of H3K4 di- and tri-methyl marks. Lid is related to the mammalian JARID family members, and although no equivalent protein has been described in mammals, LSD1, an unrelated protein can demethylate H3K4me2 (Lee et al., 2007b; Shi et al., 2004). The identification of PcG and TrxG proteins which can dynamically methylate and demethylate these opposing marks provides a possible mechanism for the resolution of bivalent domains upon lineage commitment (Fig. 1.1-2).

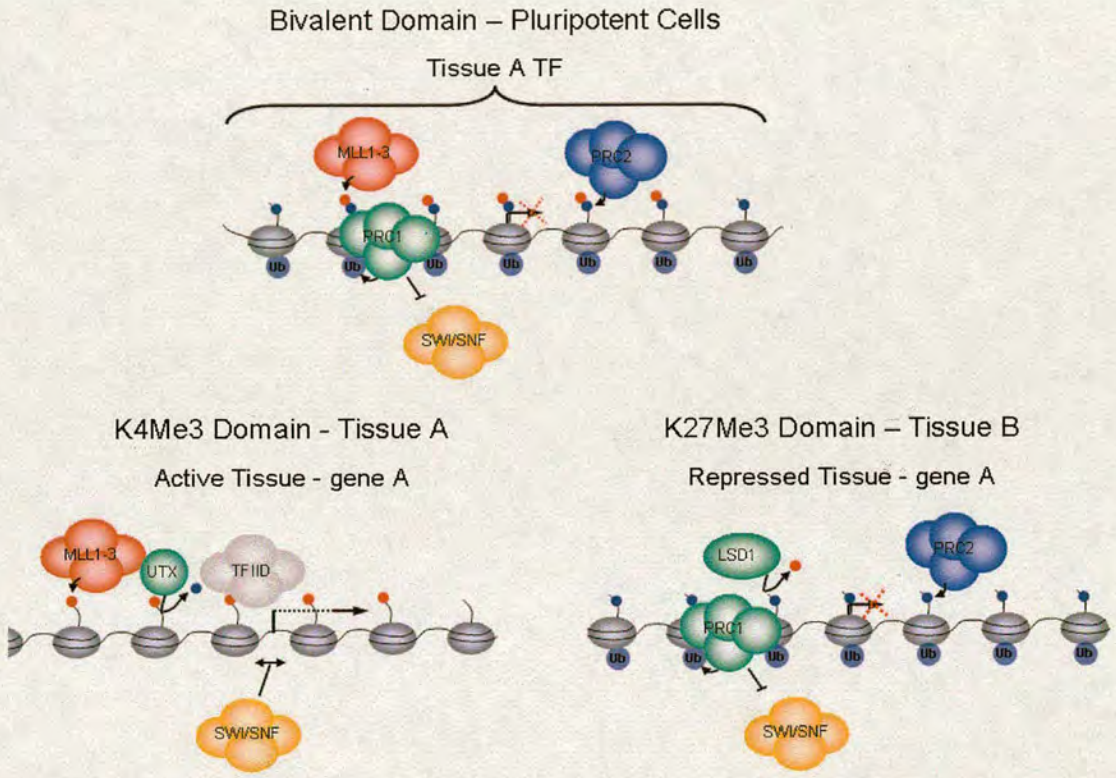


Figure 1.1-2. PcG, TrxG and Bivalent Chromatin Domains
Hypothetical schematic representing the mechanism whereby bivalent chromatin domains may be resolved by the PcG and TrxG proteins upon tissue differentiation. In ES cells the bivalent domains are maintained by the H3K27 HMTase activity of the PRC2 complex (filled blue circles) which then recruits the PRC1 complex (filled green circles) via interaction with the methyl moiety. PRC1 then ubiquitinates H2A tails at lysine 119 (data based on evidence from X inactivation; (de Napoles et al., 2004)). Chromatin remodeling by the SWI/SNF complex (filled orange ovals) is antagonized by the presence of PRC1. MLL1-3 (filled red ovals) trimethylates H3K4, but is insufficient to induce promoter activity due to the presence of the inactive marks, and lack of chromatin remodeling. Upon differentiation, the domains may then be resolved by either the demethylase activities of UTX (filled green circle) or a Lid like JARID protein (filled green ovals). In the active state SWI/SNF may afford promoter access by nucleosome remodeling and in conjunction with TFIID facilitate transcription. H3K4me3 (small red circles) and H3K27me3 (small blue circles) are indicated.

Drosophila PcG and TrxG have the ability to stably transmit memory of transcriptional activity through mitosis and meiosis (Cavalli and Paro, 1998). How do these proteins and their modifications memorise the transcriptional state following cellular division? It is

possible that semi-conservative nucleosomal replacement provides a template to perpetuate these modifications following S-phase (Martin and Zhang, 2007). Alternatively, large kilobase tracts of H3K27 methylation may facilitate the modification of newly assembled chromatin through association with PRC2 present downstream of the replication forks (Bernstein et al., 2006; Cao and Zhang, 2004; Mikkelsen et al., 2007). Another possibility is that DNA methylation status or other external factors could recruit PcG and TrxG proteins following replication. Accordingly, depletion of DNMT1 in the human U2OS cells was shown to disrupt the localisation of PRC1 subunits (Hernandez-Munoz et al., 2005). The rapid accumulation of knowledge pertaining to this field will hopefully provide mechanistic insight into the heritable nature of the PcG and TrxG systems.

1.2 DNA Methylation

The epigenetic mechanisms so far discussed involve alteration of chromatin structure via modulation of its protein components. However, DNA bases can be chemically modified directly through the addition of methyl-groups to the carbon-5 (C5) and carbon-6 (C6) positions of cytosine and adenine bases respectively. DNA methylation exists in all kingdoms of life, although the distribution of the methyl mark varies widely between species. Functionally it has been implicated in transcriptional repression, mammalian imprinting, suppression of transposition and recombination and chromatin organisation.

1.2.1 Distributive Diversity of DNA methylation

DNA Methylation in Fungi

5-Methylcytosine (5MeC) is the most prevalent modified base in eukaryotes, although not all species possess the modification. In fungi, neither of the ascomycetes *S. cervisiae* and *S. pombe* possesses DNA methylation (Proffitt et al., 1984). In contrast, the filamentous fungi *Neurospora crassa* has methyl-cytosine levels comparable to that observed in mammals (Selker et al., 2003). In *Neurospora*, DNA methylation is almost exclusively localised to relics of repetitive elements, disrupted by RIP (Repeat induced Point mutation; (Selker, 2002; Selker et al., 2003)). Consistent with this idea, ablation of DNA methylation via 5AzaC treatment has been shown to reactivate transposition (Zhou et al., 2001). The related species *Ascobolus* shows a similar methylation distribution which has been implicated in the repression of deleterious recombination events (Maloisel and Rossignol, 1998). This data is

consistent with DNA methylation providing a genome defense mechanism against parasitic DNA elements.

DNA Methylation in Plants

A broad spectrum of DNA methylation composition is observed in plant species. The *Arabidopsis* genome (120Mb) contains 6% 5MeC which is approximately quarter of the level found in the maize genome (2500Mb; (Bender, 2004)). In contrast to most eukaryotes, *Arabidopsis* contains methylation in the context of both CpG, and non CpG sequences. The methylation is distributed across repetitive sequences including telomeric, centromeric and pericentromeric chromatin as well as transposable elements (Cokus et al., 2008; Zhang et al., 2006; Zilberman et al., 2007). The majority of genes are completely unmethylated, however approximately 20-30% are heavily methylated at CpG[ii] dinucleotides which correlate with transcriptional activity (Cokus et al., 2008; Zhang et al., 2006; Zilberman et al., 2007). In contrast, the small proportion of genes which possess promoter methylation show a more tissue restricted expression profile (Zhang et al., 2006). *Met1* mutation results in the expression of pseudogenes and transposons suggesting that CpG methylation is required for transcriptional repression (Zhang et al., 2006). Similar to various fungi, this data is indicative of a central genome defense function for DNA methylation in plants.

DNA Methylation in Invertebrates

DNA methylation has been most extensively investigated in animals, which has provided a detailed picture of its genomic distribution (Tweedie et al., 1997). Invertebrates invariably contain low level methylation relative to their vertebrate counterparts (Tweedie et al., 1997). Indeed nematodes such as *Caenorhabditis elegans* lack DNA methylation and only 0.4% of ^{me}CpG (Methyl-CpG) is detected during early *Drosophila* embryogenesis (Gutierrez and Sommer, 2004; Lyko et al., 2000). The sea squirt (*Ciona intestinalls*) is a urochordate and therefore represents one of the most 'vertebrate-like' invertebrate species. It contains a mosaic pattern of DNA methylation which is reminiscent of more primitive invertebrates such as the sea urchin (Echinoderm) (Simmen et al., 1999; Suzuki et al., 2007). Methylation is present in distinct compartments and preferentially localizes to more than half of sea squirt genes, whilst showing no preference for repetitive elements (Simmen et al., 1999; Suzuki et al., 2007). Moreover, the methylated genes were found to be moderately expressed

[ii] CpG methylation is catalysed by the Met1 methyltransferase in *Arabidopsis* which is homologous to mammalian DNMT1 (Chan, S.W., Henderson, I.R., and Jacobsen, S.E. (2005). Gardening the genome: DNA methylation in Arabidopsis thaliana. Nature reviews *6*, 351-360.).

housekeeping genes, equivalent to a similar distribution in *Arabidopsis* (Suzuki et al., 2007; Zilberman et al., 2007). These findings illustrate that invertebrate genomes are divided up into compartments with respect to DNA methylation, and that transition boundaries between these domains can occur very sharply as illustrated in *Ciona* (Suzuki et al., 2007).

DNA Methylation in Vertebrates

Interestingly the development of a backbone coincides with a gross transition from fractional to global CpG methylation at the vertebrate invertebrate boundary (Tweedie et al., 1997). Analysis of vertebrate methylation indicated that even the genomes of jawless fish (*Agnathans*), the most primitive vertebrates, are heavily methylated (Tweedie et al., 1997). Whilst invertebrate gene methylation is largely restricted to those with housekeeping function[iii], all vertebrate genes are methylated throughout the majority of their length (Suzuki et al., 2007; Tweedie et al., 1997). Moreover, ribosomal DNA (rDNA) sequences have acquired mosaic methylation in vertebrates consistent with intragenic methylation (Bird, 1980; Tweedie et al., 1997). It is unclear whether genome defense is a major function of vertebrate DNA methylation as 5MeC is distributed globally and largely independent of specific sequence. However, global hypomethylation associated with human neoplasia has been shown to reactivate certain transposable elements suggesting that this function is important, but may be partially redundant with additional repressive mechanisms (Fukasawa, 2005; Rodriguez et al., 2006). It has been posited that global methylation arose due to the increased gene content and genomic complexity of the vertebrate lineage (Bird, 1995). Therefore gross alteration of DNA methylation may have resulted in the acquisition of specialized functions to combat the increased transcriptional burden associated with expanded genome size (Colot and Rossignol, 1999).

1.2.2 DNA Methylation in Mammals

Genetic analysis has shown that DNA methylation is essential for mammalian development (Li et al., 1992; Okano et al., 1999). Somatic methylation is restricted to cytosines in the context of the palindromic dinucleotide CpG and, as for all vertebrates, is heavily methylated at the majority of these sites (approximately 70% of CpGs; (Ehrlich et al., 1982)). The genome can be categorised into two distinct fractions with respect to DNA methylation and sequence composition. The major fraction (~98%) is methylated at almost all CpGs sites and, due to the process of spontaneous deamination, is CpG deficient (~0.21

[iii] Housekeeping genes, are characterised by ubiquitous expression in all or most cell types, indicative of a primary role in cellular regulation.

observed/expected CpG frequency; (Bird, 1980)). This fraction includes both intragenic regions, transposable elements and endogenous repeat sequences, although it is somewhat depleted for regulatory sequence elements (Eckhardt et al., 2006; Rollins et al., 2006; Weber et al., 2005). The minor fraction (<2%), is generally devoid of methylation and therefore shows little CpG suppression (Bird, 2002; Bird et al., 1995). This fraction represents short stretches of DNA sequence which are termed CpG Islands (see section 1.3; (Bird et al., 1985)).

Methylation patterns are established *de novo*, during gametogeneisis, embryogenesis and cellular differentiation (Reik, 2007; Reik et al., 2001; Weber and Schubeler, 2007). Consequently, there is a requirement for global demethylation to remove somatic methylation, and subsequent re-methylation to establish the appropriate genomic distribution during development. Reprogramming is required to erase epigenetic abnormalities which would otherwise be perpetuated across subsequent generations. Once established, 5MeC patterns are maintained through a postsynthetic semi-conservative copying mechanism (see section 1.2.4).

Demethylation: Setting and Resetting the Methylation Stage

In mammalian development there are two distinct phases in which DNA methylation is stripped from the genome. During gametogenesis, the primordial germ cells transit to the genital ridge prior to the formation of the gonads. During embryonic days 8-12.5 (E8-12.5), postmigratory germ cells undergo global demethylation including single copy genes, repetitive elements and parentally imprinted sequences (Hajkova et al., 2002; Kafri et al., 1992; Lee et al., 2002; Maatouk et al., 2006). This temporal demethylation process is equivalent in both male and female embryos and is required to reprogram methylation patterns established in the preimplantation embryo. It is unclear whether demethylation at this stage is an active or passive process[iv]. A recent study identified extensive alterations in chromatin composition and modification both prior to and during the phase of demethylation (Hajkova et al., 2008). Global reduction of histone modifications during this period suggests that an active demethylation may proceed through a repair mechanism equivalent to that observed in plants (Hajkova et al., 2008). Prospermatogonia initiate *de novo* methylation at

[iv] There are two potential mechanisms whereby DNA methylation can be removed from DNA. Transient demethylation implies that cell division procedes in the absence of maintenance methylation at the replication forks which results in a 50% decrease in methylation level per cell division. Alternatively, active demethylation requires an enzymatic activity to catalyse the removal of the methyl moiety from the cytosine base. Active demethylation is independent of cell division and can result in extensive DNA hypomethylation very rapidly.

approximately E15-16 (Davis et al., 1999; Davis et al., 2000; Kafri et al., 1992; Lees-Murdock et al., 2003; Ueda et al., 2000). Developing oocytes re-acquire DNA methylation at a later stage, with complete re-methylation being achieved post partum (Chaillet et al., 1991). Temporal differences in *de novo* methylation during gamete maturation may reflect the substantial mechanistic differences between spermatogenesis and oogenesis. Reprogramming is required to activate pluripotency factors and to reinstate parent specific imprints which are essential for development (Maatouk et al., 2006). The global dynamics of DNA methylation during germ cell development are summarised in Fig. 1.2-1a.

A second wave of DNA demethylation occurs during the first 7-cell divisions of embryogenesis prior to the formation of the blastocyst. In contrast to gametic demethylation, the temporal and mechanistic events are distinct for the maternal and paternal genomes. Upon fertilization, DNA derived from sperm is relatively methylated in comparison to that of the oocyte (Kafri et al., 1992; Mayer et al., 2000; Monk et al., 1987). The paternal pronucleus is then rapidly demethylated following zygote formation. This was elegantly illustrated in two studies through immunostaining with an antibody specific for 5MeC (Mayer et al., 2000; Santos et al., 2002). Demethylation was shown to occur rapidly, with extensive depletion of 5MeC by 4-6 hours post fertilization (Mayer et al., 2000; Oswald et al., 2000; Santos et al., 2002). This suggests demethylation proceeds via an active process as it occurs in the absence of cell division, although no enzymatic activity has thus far been identified (Santos et al., 2002). In contrast to this active process, the female pronucleus undergoes transient demethylation during the 4-6 cell division stage (Mayer et al., 2000; Santos et al., 2002). This passive demethylation is concomitant with a reduction in methyl-transferase activity at this stage suggesting that it occurs due to lack of maintenance methylation following DNA synthesis (Monk et al., 1991). Global *de novo* methylation of both parental genomes is observed by the blastocyst stage of embryogenesis (Kafri et al., 1992; Santos et al., 2002). This methylation is established by the *de novo* methyltransferases which are highly expressed during this phase of embryogenesis (see section 1.2.4).

There are sequences which avoid demethylation, including *IAP* elements and parentally imprinted sequences (Howlett and Reik, 1991; Huntriss et al., 1998; Tremblay et al., 1997). A recent study identified Stella, a maternal specific protein factor which localizes to DNA. Genetic ablation of this protein is embryonically lethal and leads to hypomethylation at various imprinted regions and *IAP* elements, consistent with a role in maintenance of site specific methylation (Nakamura et al., 2007). Mechanistically, protection could be afforded

by steric hindrance of the demethylation machinery. Alternatively, maintenance of methylation at specific sites in the maternal genome may be facilitated by recruitment of Dnmt3L and associated *de novo* methyltransferases. This mechanism is supported by the observation that Dnmt3L deficiency results in the loss of parental specific methylation imprints in murine cells (Bourc'his et al., 2001; Hata et al., 2006; Hata et al., 2002; Webster et al., 2005). Reprogramming of methylation during early embryonic development is summarised in Fig. 1.2-1b.

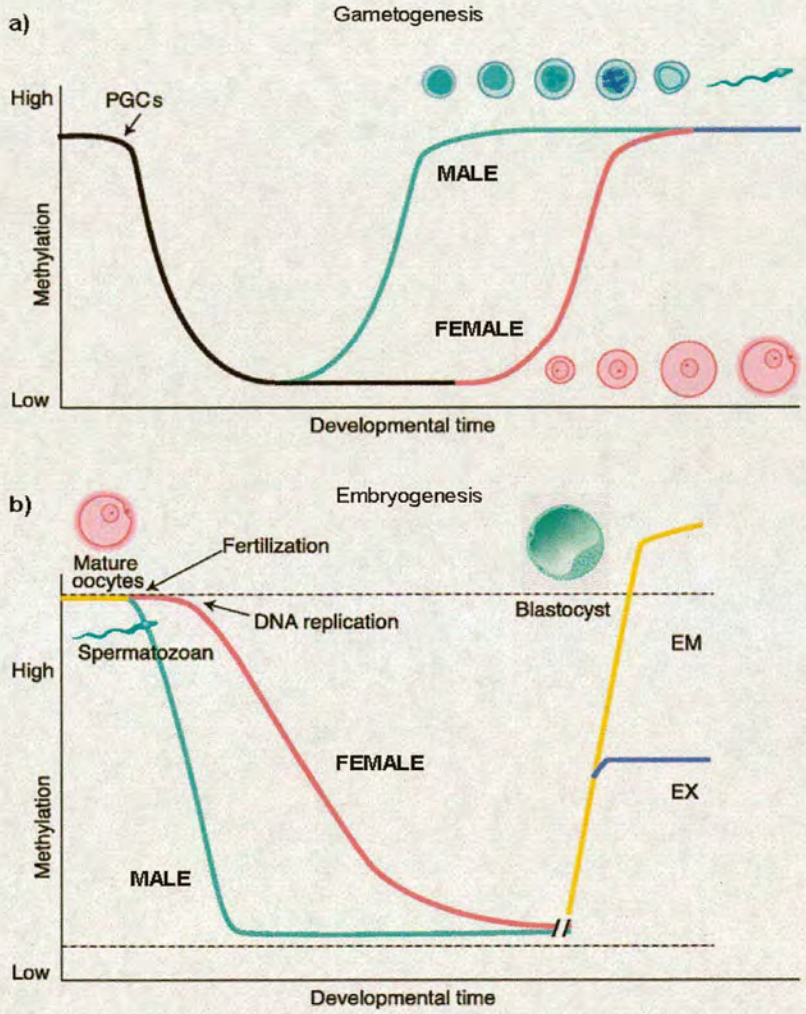


Figure 1.2-1. Global DNA Methylation Dynamics During Mammalian Development
a) Schematic representation of methylation levels during gametogeneisis. Both paternal and maternal primordial germ cells (PGCs) are heavily methylated following *de novo* methylation in the preimplantation embryo. Subsequently, global DNA demethylation occurs at approximately E11.5 which is then followed by *de novo* methylation of the paternal PGCs at approximately E13.5. *De novo* methylation occurs later in the female germ line such that it is not complete until after birth. **b)** Schematic representation of methylation levels during embryogenesis. Upon fertilization, the paternal pronucleus is actively demethylated within 4-6 hours. Subsequently, the maternal genome is passively demethylated during cleavage stage cell divisions 4-6. Nuclear DNA from cells of the extraembryonic (EX) and embryonic (EM) lineages are then *de novo* methylated by the formation of the blastocyst, although methylation levels are lower in the extraembryonic cells. Figure adapted from (Reik et al., 2001).

The mechanism of active DNA demethylation in mammalian cells remains elusive. Biochemical analysis has thus far been precluded by the limited quantities of protein available at these early developmental stages. However, in *Arabidopsis* site specific demethylation is achieved by the DNA glycosylase DEMETER, ROS1 and DML (Morales-Ruiz et al., 2006; Penterman et al., 2007). Moreover, there have been studies which claim to have identified equivalent activities in mammalian systems, although confirmatory evidence has not been forthcoming (Bhattacharya et al., 1999; Ng et al., 1999). In murine cells demethylation of a regulatory site in the *interleukin-2* (*IL-2*) promoter region, is required for activation of transcription during T-cell differentiation (Bruniquel and Schwartz, 2003). Furthermore, two recent studies indicate that methylation levels vary periodically at the *pS2/TTF1* promoter upon oestrogen activation. Intriguingly, it has been proposed that methyl groups are removed by deamination facilitated by the DNMTs themselves (Kangaspeska et al., 2008; Metivier et al., 2008). The role of DNA demethylation during embryogenesis and the potential for more specific functions in terminally differentiated cells, will likely fuel future investigation into this enigmatic process.

1.2.3 Function of Mammalian DNA Methylation

Mice deficient for the major DNMTs have reduced DNA methylation levels and die during embryogenesis (Li et al., 1992; Okano et al., 1999). Also global hypomethylation and concomitant gross chromosomal rearrangements are common features of neoplasia (Jones, 2002). The severe consequence caused by perturbations in correct genome-wide methylation levels illustrates the functional importance of DNA methylation in mammalian cells.

Transcriptional Repression

One of the best characterised functions of DNA methylation is transcriptional repression. Initial evidence supporting this mechanism came from reporter assays carried out in the oocytes of the amphibian *Xenopus laevis*. Expression of *Ade2* driven from the *E2a* promoter region of adenovirus was extinguished upon artificial introduction of methyl-groups within this sequence (Vardimon et al., 1982). An equivalent result was observed when the coding sequence of *APRT* was transfected into mouse L-cells (Stein et al., 1982). More recently, global methylation analysis using DNA microarrays identified a panel of gene promoters which associate with methylated CGIs. These methylated islands were found to be devoid of RNA polymerase II, and therefore transcriptionally silent (Weber et al., 2007). However these observations do not preclude the possibility that DNA methylation is a secondary event

during transcriptional silencing. Several independent studies have described the initiation of transcriptional repression prior to the accumulation of DNA methylation (Gautsch and Wilson, 1983; Lock et al., 1987). A striking example of this occurs during X-inactivation. A coordinated process leads to transcriptional repression of the majority of genes on one of the two X chromosomes. Despite the requirement for DNA methylation in this process, *de novo* methylation does not accumulate at gene promoters until after transcriptional repression is completed (Csankovszki et al., 2001; Gilbert and Sharp, 1999; Lock et al., 1987; Norris et al., 1991; Tribioli et al., 1992; Wutz and Jaenisch, 2000; Yen et al., 1984). These findings indicate that DNA methylation may facilitate stable transcriptional silencing, rather than being directly involved in its establishment. However, the application of the DNA methyltransferase inhibitor 5AzaC was shown to reactivate certain genes that were previously methylated on the inactive X chromosome (Xi; (Mohandas et al., 1981)). The role in transcriptional repression is further complicated by the observation that tissue-specific gene promoters containing low density CpG methylation can initiate transcription efficiently (Weber et al., 2007).

TFs bind to regulatory elements and recruit RNA polymerase II (RNA polII) to initiate transcription (Fig. 1.2-2a; (Ptashne, 2005)). In light of this fact, there are two tenable mechanisms whereby methylation-dependent transcriptional repression could be achieved. The first is via direct steric hindrance of TF binding where methyl-groups located in the major groove antagonize association with consensus sites (Fig. 1.2-2b). Consistently, methylation of the recognition sites of the transcription factors E2F1, MLTF and ETS have been shown to inhibit DNA binding concomitant with transcriptional repression (Campanero et al., 2000; Gaston and Fried, 1995; Watt and Molloy, 1988). Further support was provided by investigation of the chromatin boundary element CTCF, which functions as an insulator, separating promoters from distal enhancer elements (Filippova, 2008). CTCF binding is inhibited by DNA methylation which is known to function in imprinted transcriptional regulation of the *Igf2r* locus (Bell and Felsenfeld, 2000; Szabo et al., 2000). However, binding of various ubiquitous transcription factors including Sp1 are unaffected by DNA methylation (Harrington et al., 1988). Accordingly, certain mammalian genes can be repressed by DNA methylation, even when transcription factor occupancy is unaffected (Becker et al., 1987). Therefore, the steric hindrance model alone is insufficient to account for methylation-dependent transcriptional repression. Consistently, recruitment of transcriptional co-repressors has been shown to indirectly mediate gene silencing. Initial support for this model was provided by the purification a repressor protein from mammalian

nuclear extracts which had specific affinity for clusters of ^{me}CpG sites (Boyes and Bird, 1991; Meehan et al., 1989). Indirect gene silencing was suggested by the ability to titrate out this repressive activity using non-specific methylated competitor DNA in HeLa nuclear extracts (Boyes and Bird, 1991). Subsequently, a family of related ^{Me}CpG Binding Domain (MBDs) co-repressors were identified, several of which could invoke transcriptional repression by facilitating the formation of a repressive heterochromatin environment (Fig. 1.2-2c; (Cross et al., 1997b; Hendrich and Bird, 1998; Lewis et al., 1992)). MBD proteins will be discussed in section 1.2.7.

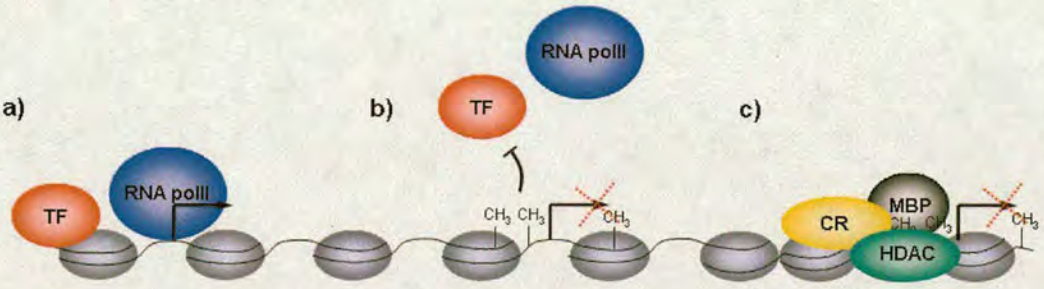


Figure 1.2-2. DNA methylation mediated transcriptional repression
a) Transcription initiation from a protein coding gene promoter with RNA polymerase II (RNA polII; filled blue circle) and sequence specific transcription factor (TF; filled red circle) bound proximal to the transcription start site (arrowed). **b)** Transcriptional repression resulting from steric hindrance of transcription factor binding by CpG methylation in proximity to the promoter. **c)** Transcriptional repression by methyl-binding protein (MBP; filed grey circle) recruitment and the association of histone deacetylase (HDAC; filled green circle) and chromatin remodellers (CR; filled yellow circles). Chromatin remodeling is indicated by the reduced separation between nucleosomes (filled light blue circle). Transcriptional repression (dashed red cross) and methylated CpGs (CH_3) are indicated.

Silencing Cryptic Promoter Elements

In other species there is evidence to suggest that gene body methylation results in the inhibition of transcriptional elongation, leading to aborted transcripts (Barry et al., 1993). However, mammalian genes are heavily methylated throughout their coding sequence whilst being transcriptionally permissive suggesting that intragenic methylation does not inhibit transcriptional elongation in mammalian cells (although there is evidence to suggest that high density methylation can reduce its efficiency; (Hellman and Chess, 2007; Jones, 1999; Lorincz et al., 2004)). Interestingly, low density methylation was shown to be sufficient to repress transcription from a weak promoter, but not from a more potent one containing an enhancer element (Boyes and Bird, 1992). Taken together this data suggests that DNA methylation suppresses transcription initiation in a ^{me}CpG density dependent fashion. This suggests that gene body methylation provides a mechanism to 'dampen' spurious transcription by silencing cryptic promoter elements within intragenic regions (Bird, 1995; Suzuki and Bird, 2008). Accordingly, DNA methylation in plants and invertebrates

frequently associates with the bodies of moderately transcribed genes which may be more susceptible to spurious internal transcription (Suzuki et al., 2007; Zilberman et al., 2007). This proposed mechanism is analogous to H3K36 methylation of intragenic transcribed regions in yeast, which is deposited following the passage of RNA polymerase to prevent internal transcription initiation events (Carrozza et al., 2005).

Antisense Transcriptional Repression

An alternative function of intragenic methylation is to silence antisense transcripts which can negatively regulate the sense transcript. Indeed such ncRNAs have been reported to repress transcription of homologous sequences, some of which are transcribed in a methylation sensitive fashion (Panning and Jaenisch, 1996; Rinn et al., 2007; Sleutels et al., 2002; Wutz et al., 1997). Whether this represents a general silencing mechanism is unclear, but two lines of evidence support this possibility. 1) Global transcription studies have detected a large number of functionally unknown ncRNAs which may play a role in such a mechanism (Bertone et al., 2004; Birney et al., 2007). 2) Double stranded RNA arising from transcription of both DNA strands can invoke RNA mediated transcriptional silencing (discussed in section 1.2.5; reviewed in (Yang and Kuroda, 2007)).

Genome Defense

Despite a major function in transcriptional repression, the majority of genomic ^{me}CpG is localised outwith regulatory elements (Rollins et al., 2006). Furthermore, the global distribution of DNA methylation across the genome suggests that it may provide additional functions. DNA methylation has been shown to maintain genome integrity in various non-mammalian species through the suppression of homologous recombination and transposition of parasitic sequences (Barry et al., 1993; Maloisel and Rossignol, 1998; Selker, 2002; Selker et al., 2003). There is evidence which suggests that DNA methylation may provide an equivalent stabilizing function in mammalian cells.

CpG deficiency results from the fixation of C to T transition mutations arising due to germ line methylation at these sites (Bird, 1980). An analysis of DNA sequence composition determined that various transposable elements were extensively CpG-depleted indicating that these sequences are hypermethylated in the germline (Rollins et al., 2006). Association of methylation with these potentially deleterious sequences suggests a possible role in genome defense. Deficiency of Dnmt3L, a factor known to facilitate *de novo* DNA methylation, results in extensive hypomethylation and concomitant expression of dispersed repeat

sequences during spermatogenesis (Bourc'his and Bestor, 2004; Gowher et al., 2005; Hata et al., 2002)}. Deficient spermatocytes showed extensive meiotic chromosomal abnormalities which precluded the formation of mature sperm cells (Bourc'his and Bestor, 2004; Webster et al., 2005). Similarly, Dnmt1 deficient mouse embryos display a significant reduction in methylation levels and elevated expression of various classes of *IAP* proviral sequences (Walsh et al., 1998). Targeted mutation of the active site of Dnmt3b in MEFs resulted in global hypomethylation, chromosomal instability and aneuploidy, the latter of which is indicative of defects in mitotic segregation and chromosomal rearrangements (Dodge et al., 2005).

Maintenance of Genomic Integrity

Global hypomethylation and chromosomal abnormalities are hallmarks of neoplastic cells (Fukasawa, 2005; Rodriguez et al., 2006; Weber et al., 2005). Chromosomal segments with an elevated incidence of rearrangements were generally found to be more hypomethylated relative to adjacent parenchyma from matched colorectal biopsy samples (Rodriguez et al., 2006). In the human cancer cell line HCT116, disruption of the catalytic domain of human DNMT1 resulted in cell cycle arrest. It was proposed that this was due to activation of the damage-induced G2/M checkpoint in response to double stand breaks in the DNA. Gross abnormalities occurred despite a modest depletion of methylation, suggesting that subtle disruption is sufficient to cause the observed chromosomal lesions (Chen et al., 2007). Moreover, an independent study identified extensive chromosomal rearrangements in HCT116 cells deficient for both DNMT1 and DNMT3b (Karpf and Matsui, 2005). Dnmt1 hypomorphic mice crossed onto a tumour prone background showed elevated loss of heterozygosity, which is indicative of increased homologous recombination or the production of unresolved double strand breaks (Eden et al., 2003). These findings suggest that global hypomethylation may promote tumorigenesis by facilitating chromosomal disruption. Further evidence for a genome integrity function comes from patients afflicted with the genetic disorder ICF (immunodeficiency, centromeric heterochromatin and facial anomalies). ICF is characterised by chromosomal segregation defects which are associated with hypomethylated pericentric satellite repeat sequences (Ehrlich, 2003). Consistent with the hypomethylated phenotype, analysis of five ICF patients identified mutations in the coding sequence of the *de novo* methyl transferase DNMT3B (Xu et al., 1999). These data suggests that global DNA methylation in vertebrates and more specifically mammalian species may have a function in maintaining the structural integrity of the genome.

1.2.4 DNMTs: Methylating the genome

A family of enzymes collectively referred to as the DNMTs is responsible for establishing and propagating genomic methylation patterns. DNMTs are divided into two classes with respect to template specificity. The *de novo* DNMTs bestow methylation upon unmodified DNA, whilst a second class propagates methylation by copying the pattern postreplicatively.

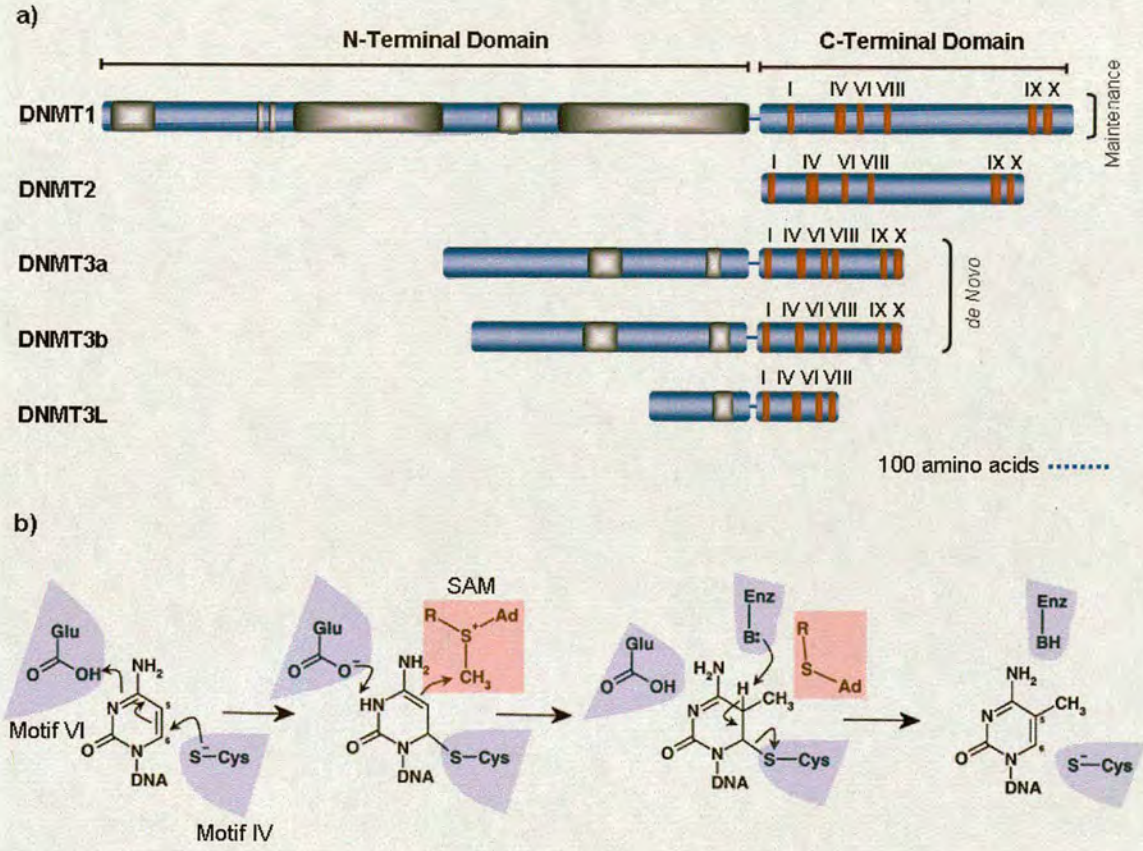


Figure 1.2-3. Structure and Catalytic mechanism of the DNMTs
a) Schematic representation of the domain structure of the five main mammalian DNMTs. The carboxy terminal domain contains the conserved catalytic motifs (roman numerals; red bars) whilst the variable amino terminal region contains (grey bars) domains required for intracellular delivery and allosteric regulation of the catalytic C terminus (adapted from (Hermann et al., 2004)). **b)** Transfer of a methyl group from the donor molecule SAM (red shading) to the C5 position of a cytosine residue. The catalytic mechanism is mediated by the DNMT C terminal domain as determined for bacterial methyltransferase enzyme *M.*Hha*I*. Major enzyme sequence motifs mediating the reaction are indicated (light blue shading; adapted from (Jeltsch, 2002).

All 5MeC DNMTs thus far characterised share a highly conserved catalytic domain which was identified by investigating bacterial methyltransferases (Klimasauskas et al., 1994; Reinisch et al., 1995). This domain comprises 10 highly conserved peptide motifs located in the carboxy terminal domain (Fig. 1.2-3a; (Hermann et al., 2004; Jeltsch, 2002)). These motifs fold into a characteristic β-sheet flanked by one or more α-helices to create both the active site and binding interface for SAM (S-adenosyl-L-Methionine) the methyl donor

molecule. Functional differences between the DNMTs occur due to specific motifs encoded in the variable amino-terminal domain (Klimasauskas et al., 1994; Reinisch et al., 1995).

The methyl group is enzymatically transferred from the SAM to the C5 position of the cytosine pyrimidine ring. Initially, the DNMT disrupts C/G base pairing and flips the cytosine base out of the DNA helix to facilitate catalysis. SAM is bound to motifs I, III and X of the catalytic domain and is destabilized by interaction with a conserved cysteine residue which makes it receptive to nucleophilic attack. Nucleophilic attack at the C6 position by a thiol group in motif IV is stabilised by protonation of the pyrimidine ring by motifs VI and VIII. Electrophilic attack by the activated Carbon 5 position followed by deprotonation resolves the reaction intermediates with the release of S-adenosyl-L-homocysteine (Fig. 1.2-3b; (Hermann et al., 2004; Jeltsch, 2002)).

DNMT1

DNMT1 is the maintenance methyltransferase required to faithfully propagate patterns of DNA methylation following DNA replication. DNMT1 processively copies the methylation pattern onto the newly synthesized DNA strands, by preferentially methylating hemimethylated DNA (Bestor, 1992; Yoder et al., 1997). Coordinated replication and DNA methylation is achieved by association of DNMT1 with PCNA (proliferating cell nuclear antigen) and the replication forks during S-phase. Interaction studies determined that localisation was facilitated by a charged region and a PCNA interaction motif (Fig 1.2-4; (Chuang et al., 1997; Leonhardt et al., 1992)). The preference for hemimethylated DNA is due to the amino terminal domain as deletion of this region abolishes this specificity and increases DNMT1's intrinsic *de novo* activity (Fig. 1.2-4; (Bestor, 1992; Chuang et al., 1997; Yoder et al., 1997). It has been suggested that this binding specificity is encoded by the CXXC domain within the N terminal domain, which has been shown to bind DNA in a methylation dependent manner in other proteins (Bestor, 1992; Jorgensen et al., 2004; Lee et al., 2001). The amino terminal domain also contains a polybromo domain which consists of two BAH motifs, and is implicated in chromatin association (Fig. 1.2-4; (Hermann et al., 2004)). Targeting to the nucleus is achieved through an NLS (Nuclear localisation signal; Fig. 1.2-4), however, during early embryonic development, Dnmt1 is excluded from the nucleus. This is thought to facilitate the passive demethylation of the female pronucleus in the preimplantation embryo. During the wave of global demethylation in gametogenesis, certain imprinted DNA loci maintain DNA methylation. Dnmt1o is a DNMT1 isoform specific to preimplantation embryos and developing oocytes and is believed to maintain

methylation at these target sequences (Ratnam et al., 2002). Consistent with its function, Dnmt1 is ubiquitously expressed in all replicating somatic cells.

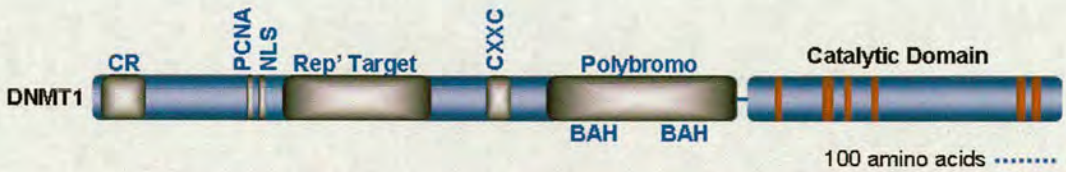


Figure 1.2-4. Dnmt1 domain structures
Schematic representation of the domain structure of DNMT1. The C terminal domain contains the conserved catalytic motifs (red bars). All charcaterised amino terminal domains are indicated (grey bars). The charged region (CG), PCNA binding site (PCNA), Nuclear localisation signal (NLS), replication foci targeting region (Rep' Target), CXXC domain and the polybromo domain consisting of 2 BAH motifs (adapted from (Hermann et al., 2004)).

DNMT1 deficiency in mouse is embryonic lethal, terminating shortly after gastrulation and is coincident with a 70% reduction in DNA methylation levels (Li et al., 1992). The viability of cultured ES cells despite a comparable loss of methylation suggests that the mouse phenotype is caused by an inability to complete the developmental programme (Li et al., 1992). Despite the global loss of DNA methylation, the ability to *de novo* methylate an integrated proviral sequence was unaffected in *DNMT1* deficient ES cells (Lei et al., 1996).

DNMT2

Mammalian DNMT2 was identified as a homolog of yeast *pmt1* by homology to the carboxy terminal domain (Okano et al., 1998b; Yoder and Bestor, 1998). DNMT2 possesses an intact catalytic domain; however preliminary analysis indicated that it was unable to methylate DNA *in vitro* (Okano et al., 1998b). DNMT2 deficiency, both in ES cells and transgenic mice, resulted in normal physiology and did not disrupt the ability to methylate a newly integrated retroviral sequence (Lei et al., 1996; Okano et al., 1998b). Taken together, these results suggested that DNMT2 provides neither *de novo* nor maintenance methylase activities *in vivo*. In *Drosophila*, discovery of low level DNA methylation during early embryogenesis led to the identification of a single *DNMT* coding sequence homologous to mammalian *DNMT2* (Kunert et al., 2003; Lyko et al., 2000). Depletion of the endogenous transcript by RNAi resulted in complete loss of all DNA methylation, indicating that it provides the sole methyltransferase activity in *Drosophila* (Kunert et al., 2003). Subsequently, a refined HPLC/TLC technique determined that human DNMT2 could provide residual methylase activity, but with more restricted sequence specificity than that identified for related DNMTs (Hermann et al., 2003). Recent investigation has suggested that DNMT2 might target methylation to specific tRNAs (Goll et al., 2006). Therefore, it is

unclear whether DNMT2 represents a functionless evolutionary relic or plays a more discrete role in selective methylation in mammalian cells.

DNMT3a and b

DNMT3a and b were identified as the *de novo* methyltransferases required to establish the initial methylation patterns during gametogenesis and embryogenesis. Neither protein has a significant preference for hemimethylated DNA templates implying that DNMT3s do not function as maintenance methyltransferases (Gowher and Jeltsch, 2001; Okano et al., 1999). Furthermore the expression of both enzymes is largely restricted to embryonic tissues (Okano et al., 1998a). Depletion of the Dnmt3s reduces DNA methylation levels despite the fact that Dnmt1 has been shown to possess *de novo* methyltransferase activity (Bestor, 1992; Okano et al., 1999; Yoder et al., 1997). Both *de novo* DNMTs have a PWWP domain and PHD finger domains which are involved in transcriptional regulation and chromatin association (Fig. 1.2-5; see section 1.2.5). Interestingly, both Dnmts were found to methylate cytosines in contexts outwith CpG dinucleotides (Gowher and Jeltsch, 2001; Ramsahoye et al., 2000; Suetake et al., 2003). This is consistent with elevated non-CpG cytosine methylation levels in ES cells where the *de novo* Dnmts are highly expressed (Gowher and Jeltsch, 2001; Ramsahoye et al., 2000; Suetake et al., 2003). However it is unclear if these modifications have any biological relevance as they are absent in somatic cells due to lack of maintenance by Dnmt1 (Ramsahoye et al., 2000).

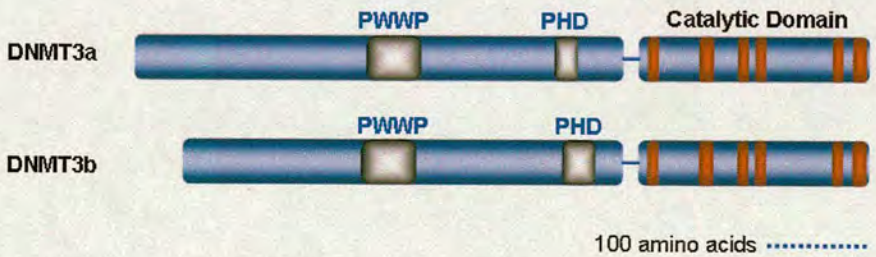


Figure 1.2-5. Dnmt3a and b domain structures
Schematic representation of the domain structure of the *de novo* DNA methyltransferases. The carboxy terminal domain contains the conserved catalytic motifs (red bars). The N terminal domain of both proteins encodes both PWWP and PHD domains (adapted from (Hermann et al., 2004)).

Sequence similarity and domain structure conservation between the two enzymes, suggests overlapping functionality *in vivo (Okano et al., 1998a)*. Consistent with this idea, both enzymes were shown to redundantly methylate an inserted retroviral sequence and *AluI* elements *in vivo* (Lei et al., 1996; Okano et al., 1999). The increased phenotypic severity of mice deficient for both *de novo* DNMTs suggests a synergistic effect (Okano et al., 1999).

However, disparate phenotypes of the individual knock-out mice indicate that the two proteins are not entirely functionally equivalent (Okano et al., 1999). Mice deficient for Dnmt3a are phenotypically normal at birth but become runted and die at approximately 4 weeks post-partum (Okano et al., 1999). Molecular analysis indicated that Dnmt3a is required for methylation of the major satellite repeats and for establishing maternal imprints through interaction with Dnmt3L (Chen et al., 2003b; Hata et al., 2002). Alternatively ablation of Dnmt3b activity, results in embryonic lethality at approximately 15.5dpc (days post coitum), and shows hypomethylation of specific minor satellite repeats (Chen et al., 2003b; Okano et al., 1999). The human genetic disorder ICF is associated with perturbations in the *DNMT3b* coding sequence, resulting in hypomethylation of specific endogenous repeats and promoter CGIs on the female Xi chromosome (Hansen et al., 2000; Okano et al., 1999; Wijmenga et al., 1998). The etiology of this disorder, in conjunction with the Dnmt3b deficient phenotype in mouse, suggests an essential role in the methylation of certain endogenous repeat sequences including the minor satellite repeat (Miniou et al., 1997; Okano et al., 1999). Another fundamental difference between the *de novo* Dnmts is the nature of their enzymatic function. Dnmt3b is processive and can track along DNA to methylate multiple CpG sites. In contrast, Dnmt3a acts distributively and appears to be targeted to discrete genomic loci (Gowher and Jeltsch, 2001; Kim et al., 2002).

DNMT3L

DNMT3L resembles the carboxy terminus of the *de novo* methyltransferases, but lacks essential sequence motifs rendering it catalytically inactive (Bourc'his et al., 2001; Hata et al., 2002). Despite this, genetic analysis has shown it to be essential for spermatogenesis and the methylation of maternally imprinted DMRs (Differentially methylated regions). Deficient male mice are infertile which is likely due to defects in sperm maturation (Bourc'his et al., 2001; Hata et al., 2006; Hata et al., 2002). Dnmt3L is temporally and spatially expressed during gametogenesis suggesting that it may be involved in the establishment of methylation patterns. Indeed the timing coincides with global methylation of PGCs and the generation of imprinted signals (Sakai et al., 2004). Furthermore, Dnmt3a and b physically interact with a carboxy terminal region of Dnmt3L *in vivo* and this can enhance their respective activities (Gowher et al., 2005; Hata et al., 2002; Suetake et al., 2004). Interestingly, Dnmt3L was also found to bind directly to DNA *in vitro* and associate with the core histones *in vivo* (Gowher et al., 2005; Ooi et al., 2007). This suggests that it may function as a modulator of activity and provide spatial targeting of *de novo* methylation *in vivo*. Preliminary investigation indicated that Dnmt3L deficiency did not affect global

methylation levels which may indicate that it functions at discrete genomic loci (Bourc'his et al., 2001). Consistently, Dnmt3L deficiency was shown to cause hypomethylation of specific endogenous repetitive sequence elements (Hata et al., 2006; Webster et al., 2005).

It is interesting to note that whilst the categorisation of Dnmts into *de novo* and maintenance functions is grossly accurate, there is a level of cooperativity between the two systems (Kim et al., 2002; Liang et al., 2002). Dnmt1 and the *de novo* methyltransferases can physically interact *in vivo* via their amino terminal domains. This interaction results in a synergistic increase in methylation activity (Kim et al., 2002). An early study indicated that the process of methylating an exogenous sequence was intrinsically low fidelity resulting in 5% methylation errors per cell division (Pollack et al., 1980; Wigler et al., 1981). Maintenance methylation of repetitive elements is highly inefficient in the absence of the *de novo* Dnmts in mice (Liang et al., 2002). Many cancers have aberrant DNA methylation, and this is often concordant with increased *de novo* DNMT activity (Roll et al., 2008; Ting et al., 2006). Moreover, cancer cells with elevated DNMT3b levels were found to generate increased numbers of methylation errors (Linhart et al., 2007; Ushijima et al., 2005). Taken together, this suggests that the *de novo* methyltransferases cooperate with Dnmt1 to accurately maintain DNA methylation patterns postreplicatively.

1.2.5 Targeting *de novo* Methylation

The mammalian genome is heavily methylated along the majority of its length; however, there are certain sequences which are more prone to the acquisition of DNA methylation. Imprinted genes, transposable elements and certain CGI sequences have all been reported to be more frequently targeted by the methylation machinery (Bock et al., 2006; Meunier et al., 2005; Reik, 2007; Weber et al., 2007). However, the mechanism by which these sequences are preferentially targeted for *de novo* methylation is still poorly understood. Conceptually, targeting could be achieved via four distinct mechanisms. 1) Dnmts may have an intrinsic preference for specific DNA sequences or chromatin structures specified by interaction motifs (Fig. 1.2-6a). 2) Interaction with protein cofactors could direct methylation to specific genomic loci (Fig. 1.2-6b). 3) Methylation may be targeted via an RNAi mediated transcriptional repression pathway (Fig. 1.2-6c). 4) DNMTs may be occluded from certain DNA sites by steric hindrance from bound factors. These mechanisms need not be mutually exclusive and could provide a failsafe mechanism for distributing the methyl-mark *in vivo*.

Sequence Specific Targeting

There is little evidence for sequence specificity for any of the main Dnmts outwith the central CpG site. However, Dnmt3a has been reported to have slight preferential activity at CpG sites flanked by pyrimidines (Lin et al., 2002). The PWWP domain (pfam: PF00855), present in the amino terminus of both *de novo* methyl-transferases is essential for targeting to chromatin (Ge et al., 2004; Qiu et al., 2002). In plants and fungi, DNA methylation is targeted to DNA sequences which are present in multiple copies (Selker et al., 2003; Zhang et al., 2006). There is less direct evidence for such an instructive mechanism in mammalian systems. Meunier and coworkers utilised CpG transition mutations as indicators of genomic methylation levels in the germline of primates. This analysis identified transposable elements as being the most heavily methylated component of the genome (Meunier et al., 2005). It has been proposed that hemimethylated sequences produced by recombination between such endogenous repeats and newly integrated homologous sequences may provide preferential target for DNMT activity (Fig. 1.2-6a; (Bestor and Tycko, 1996)).

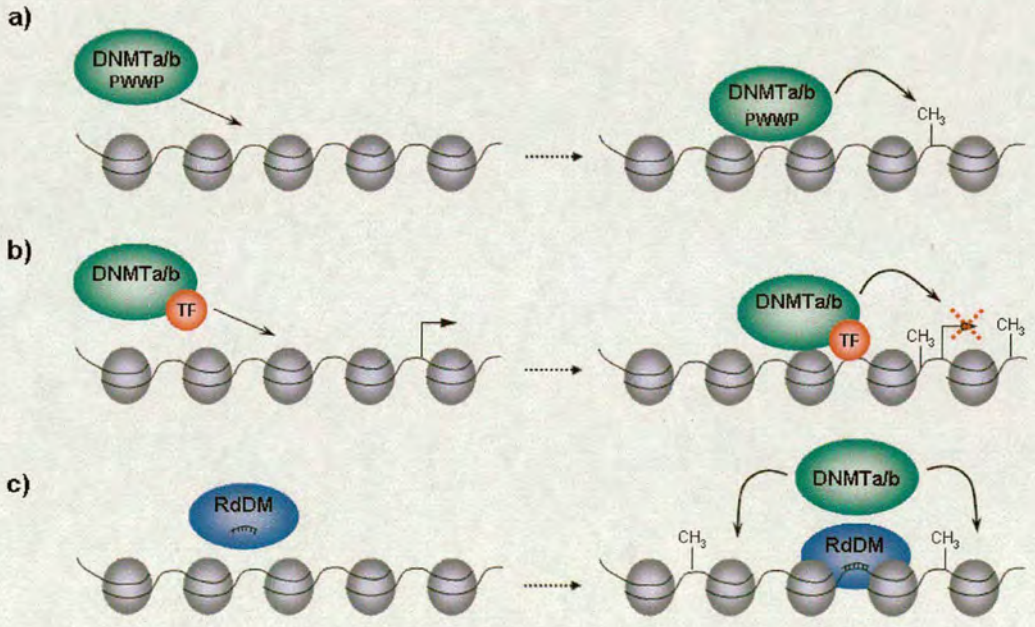


Figure 1.2-6. Mechanisms for DNA targeting.
Three possible mechanisms for DNMT targeting include; direct association (**a**), binding by sequence specific protein cofactors (**b**) and recruitment facilitated by an RNA dependent DNA Methylation (RdRM) nucleoprotein complex (c). **a**) *de novo* methyltransferase (filled green circle) can directly bind chromatin (string of blue circles) through interaction with their amino terminal PWWP domains. **b**). Sequence specific transcription factors (red circle) can direct Dnmts to gene promoters (arrowed). **c**) RNA interference pathway targets DNA in a site specific manner, through sequence recognition by a nucleoprotein complex (RdDM; light blue circle). This could recruit DNMTs to specific gene loci as has been described in plants. Figure adapted from (Klose and Bird, 2006).

Targeting by Protein Interactions

An alternative mechanism by which methylation may be targeted to specific sequences is via protein interacting partners. Dnmt3L can directly bind DNA and can associate with the four core histones *in vivo* (Gowher et al., 2005; Ooi et al., 2007). Furthermore Dnmt3L can physically associate with both *de novo* methyltransferase suggesting that it may act as a chromatin association factor (Suetake et al., 2004). Interestingly, chromatin association was found to be sensitive to H3K4 methylation, a modification associated with transcriptional activity (Ooi et al., 2007). This may indicate a mechanism whereby *de novo* DNA methylation is specifically excluded from sites of active transcription, as has been proposed for the existence of CpG Islands (see section 3.1.1). Dnmt3L may also have a role in targeting DNA methylation to imprinted gene loci. Structural determination and chromatographic purification of Dnmt3a complexed with Dnmt3L indicated the formation of a heterodimer which could then oligomerise on DNA (Jia et al., 2007). The authors proposed that this structure imposes a periodicity of active site/DNA contacts consistent with ^{me}CpG spacing observed in several imprinted genes. Dnmt3a and Dnmt3L are essential for the maintenance of methylation at imprinted DMRs and this structure may confer a preference towards these sequences (Bourc'his et al., 2001; Hata et al., 2002).

The oncogenic TF PML-RAR is found in Acute Promyelocytic Leukemia (APLs). It binds to the promoter region of *RARβ2*, a frequent target of aberrant methylation in cancer cells. Transfected PML-RAR was shown to recruit Dnmt1 and Dnmt3a to a *RARβ2* reporter and elicit transcriptional repression in a DNA methylation dependent manner (Di Croce et al., 2002). Dnmt3a was found to be targeted to the promoter of *p21Cip1* through it's interaction with Myc and the sequence specific targeting factor; Miz-1. Targeted silencing was shown to be dependent on DNA methylation (Brenner et al., 2005). NoRC (nucleolar remodeling complex) is required to silence specific rDNA genes and facilitates repression via chromatin modulation, nucleosome remodeling and DNMT activity, all of which are functionally indispensable (Santoro and Grummt, 2005; Santoro et al., 2002). Targeting of this complex was shown to be dependent on association with TTF1 which facilitated recruitment of both Dnmt1 and Dnmt3b to the promoter (Santoro and Grummt, 2005). These data suggest that, sequence specific targeting by accessory factors may be a general molecular mechanism by which Dnmts can elicit transcriptional repression (Fig. 1.2-6b).

RNA directed Targeting

It is hard to imagine a better system for sequence recognition than that facilitated by nucleotide base-pairing. RNA interference (RNAi) involves the targeting of short RNA molecules to complementary nucleotide sequences which invokes transcriptional gene silencing (TGS). This process has been extensively studied in plants, yeast and insects (Grewal and Elgin, 2007; Lippman and Martienssen, 2004; Matzke and Birchler, 2005). Short double stranded RNA molecules are generated by the activity of DICER[v], which preferentially cleaves double stranded RNA into 21-26nt oligonucleotides. The antisense strand provides sequence specificity to target a nucleoprotein complex to a complementary nucleotide sequences. This can mediate transcriptional repression, by targeted degradation of the complementary RNA transcript, or by recruiting chromatin modifying activities to the complementary DNA sequence. In plants, one consequence of the later mechanism is RNA directed DNA methylation (RdDM), which results in highly localised DNA methylation at complementary sequences (Lippman and Martienssen, 2004; Matzke and Birchler, 2005). However, the existence of an equivalent system in mammals is a contentious issue. Mammals possess much of the enzymatic machinery required for RNAi and can silence transcription by targeting synthetic short interfering RNAs (siRNAs) to endogenous gene loci (Kanellopoulou et al., 2005; Morris et al., 2004). Murine cells, deficient for Dicer have disrupted heterochromatin and are hypomethylated at certain sequences throughout the genome (Benetti et al., 2008; Kanellopoulou et al., 2005; Sinkkonen et al., 2008). Introduction of RNA molecules complementary to endogenous *EF1A* (Elongation Factor 1A), resulted in transcriptional repression and increased methylation at the locus (Morris et al., 2004). However, similar RNA targeting to the *CDH1* promoter in the human carcinoma cell line HCT116, showed reduced RNA and protein levels without a concomitant increase in DNA methylation (Ting et al., 2005). More recently, analysis of Dicer hypomorphs displayed up-regulation of a panel of genes consistent with the depletion of promoter DNA methylation (Ting et al., 2008). Two studies have indicated that an RNA mediated pathway may indirectly effect DNA methylation (Benetti et al., 2008; Sinkkonen et al., 2008). Dicer, deficiency in both mice and murine ES cells was found to result in derepression of Oct4 and stimulate telomeric recombination and elongation. These regions were also found to be aberrantly hypomethylated. Molecular analysis indicated that this was not due to DNA methylation targeting but instead the misregulation of the micro RNA cluster, miR-290. This

[v]Dicer or equivalent ribonuclease, as processing has also been described for Piwi proteins, which function in piRNA mediated transcriptional repression (Aravin, A.A., and Bourc'his, D. (2008). Small RNA guides for de novo DNA methylation in mammalian germ cells. Genes Dev *22*, 970-975..

was subsequently found, to directly repress RbL-1, a transcriptional repressor of the Dnmts. Therefore, Dicer deficiency was reported to result in DNA hypomethylation as a consequence of reduced methyltransferase activity, rather than targeted DNA methylation (Benetti et al., 2008; Sinkkonen et al., 2008). A recently discovered class of small RNAs (piRNAs; piwi interacting RNAs) have been implicated in DNA methylation targeting during mammalian spermatogenesis (Aravin and Bourc'his, 2008). At this stage it remains unclear as to whether DNA methylation targeting via an RNA mediated pathway is a primary silencing mechanism in mammalian cells. However, it is also probable that methylation can be recruited by other chromatin modifications, as an indirect response to this pathway (Fig. 1.2-6c).

1.2.6 DNMT Interactions: Routes to Repression

There is evidence to suggest that Dnmts can repress transcription directly, in a methylation independent manner (Myant and Stancheva, 2008; Rountree et al., 2000). In order to fully appreciate the role of the Dnmt proteins in transcriptional regulation, it is necessary to review some of the interactions with other nuclear components.

Several studies have identified HDACs (Histone Deacetylases) as Dnmt binding partners (Fuks et al., 2000; Myant and Stancheva, 2008; Robertson et al., 2000; Rountree et al., 2000). Transfection studies have indicated that Dnmt1 can inhibit transcription in an HDAC dependent manner (Fuks et al., 2003b; Rountree et al., 2000). Chromatographic purification of Dnmt1 confirmed these interactions and identified additional interaction partners including Rb and E2F1 (Robertson et al., 2000). Co-transfection of Dnmt1 and Rb, indicated cooperative transcriptional repression of an E2F1 target gene. This sequence specific repression was found to be independent of methyltransferase activity or *de novo* methylation (Robertson et al., 2000). The same group purified Dnmt1 complexed with the SNF2 ATP dependent chromatin remodeller (Robertson et al., 2004). SNF2 remodeling activity was shown to be essential for site specific DNA methylation and transcriptional repression, of specific rRNA genes (Santoro and Grummt, 2005). Deficiency of the putative chromatin remodeller Lsh (Lymphoid Specific helicase), results in genomic DNA hypomethylation in mice (Sun et al., 2004). A recent study has determined that Lsh can associate with Dnmt1 and Dnmt3b in vitro *and in vivo*. This interaction facilitates the association of HDACs 1 and 2, and can repress transcription when targeted to a gene promoter (Myant and Stancheva, 2008). Interestingly, DNMT1 catalytic activity is not required for this process, and targeting of this activity was not found to result in *de novo* methylation. This data suggests that

interactions with Dnmt1 and Dnmt3a may be sufficient to repress transcription, in the absence of methylation (Myant and Stancheva, 2008). Further work will be required to elucidate the role of this interaction in the hypomethylation phenotype observed in the deficient mouse model. It is interesting to postulate, that chromatin remodeling facilitates DNA methylation by exposing CpG sites which would otherwise be inaccessible due to nucleosome positioning.

Immunoprecipitation of EZH2 (see section 1.1.2: Polycomb and Trithorax proteins) from HeLa cells copurified DNMT activity from endogenous cell extracts. The interaction was mapped to the PHD domain of both *de novo* methyltransferases and to the amino terminus of DNMT1. Other components of the Prc2 complex were also found to associate with Dnmts. ChIP analysis indicated that EZH2 was required for Dnmt1 recruitment and that its depletion led to reactivation of a panel of Prc2 target genes. EZH2 depletion corresponded with a loss of Dnmt1 binding and reduced DNA methylation levels at associated gene promoters (Vire et al., 2006).

Dnmt3L has been shown to directly interact with chromatin and depletion led to a reduction in H3K9 methylation levels (Webster et al., 2005). It has previously been shown that histone methylation at this position is a prerequisite for effective DNA methylation in other model systems (Tamaru et al., 2003). Therefore this may provide a mechanism by which the *de novo* methyltransferases cooperate with other repressive histone marks to reinforce a repressed chromatin state. This may explain the inability of Dnmt3L deficient cells to complete spermatogenesis, as abnormal chromatin conformation could disrupt the synaptonemal complex and prevent the completion of meiosis (Ooi et al., 2007; Webster et al., 2005) Consistent with this scheme, there is evidence to suggest that both HP1 and SUV39h1 interact with Dnmt1 (Fuks et al., 2003a; Lehnertz et al., 2003).

DMAP1 is a transcriptional repressor which can associate with the extreme amino terminus of Dnmt1 to repress transcription, independent of catalytic activity (Rountree et al., 2000). HDAC2 is recruited to this complex late in S-phase, suggesting that this interaction may serve to remove acetyl-groups from the newly deposited histones, in a methylation dependent manner (Rountree et al., 2000). This would be in accordance with the observation that transcriptionally inert chromatin replicates late in S-phase.

These regulatory networks suggest mechanistic links between DNA methylation and other epigenetic components, including histone modifications and chromatin remodeling. However, Dnmts may play a role in DNA methylation independent transcriptional repression, by interaction with sequence specific cofactors and chromatin modulating activities. Indeed, there is data in *Xenopus* to suggest that xDNMT1 has an essential role in development, which is fully intact even in the absence of methyltransferase activity (Dunican et al., 2008). These data suggest that DNMTs have two overlapping, but distinct regulatory functions.

1.2.7 Mediators of the methyl mark

CpG methylation results in long term transcriptional repression of specific gene targets. One mechanism by which DNA methylation can be translated into transcriptional repression involves the recruitment of repressive complexes by Methyl-Binding Proteins (MBPs). Such interactions could conceptually prevent transcription, via occlusion of cognate transcription factor binding sites or direct steric hindrance of the basal transcriptional machinery. Alternatively, recruitment of nucleosome modifying activities could sequester DNA sequences into transcriptionally inert chromatin (Bird and Wolffe, 1999). This notion was supported by the observation that repression of exogenous DNA templates was delayed prior to assembly of chromatin (Kass et al., 1997). Further support for this idea came from the observation that, treatment with TSA[vi] resulted in extensive gene derepression (Yoshida et al., 1995).

MBPs were initially identified by the biochemical purification of a nuclear activity which specifically recognized methylated CpG sites, and could repress transcription *in vitro* (Lewis et al., 1992; Meehan et al., 1989). Mutational analysis of a related protein MeCP2, identified the minimal sequence motif required for DNA binding affinity and was termed the ^{me}CpG-binding domain (MBD; (Nan et al., 1993)). Databases searching led to the identification of a further 4 polypeptides bearing the conserved motif (Fig. 1.2-7; (Cross et al., 1997b; Hendrich and Bird, 1998)).

MeCP2

Methyl-CpG-binding protein 2 (MeCP2), possesses both the characteristic MBD consensus sequence and a carboxy-terminal transcriptional repressor domain (TRD; (Nan et al., 1997;

[vi] TSA is a HDAC inhibitor.

Nan et al., 1993)). The MBD domain can bind methylated CpG sites both *in vitro* and *in vivo* (Nan et al., 1993). Interestingly, preferential binding was observed for a portion of MeCP2 containing the MBD (amino acids 77-161) for a central methylated CpG site flanked by a run of [A/T] ≥4 residues (Klose et al., 2005). This selectivity seems to prevent MeCP2 from occupying methyl CpG sites usually bound by other members of the MBD family (Klose et al., 2005). Methyl-specific transcriptional repression was identified and linked to an association with a Sin3a HDAC corepressor complex (Nan et al., 1998). Furthermore, MeCP2 has been shown to associate with the histone H3K9 methyltransferase, Suvar 3-9 *in vivo* (Fuks et al., 2003b). Such interactions are likely to be temporally regulated or transient as the majority of cellular MeCP2 is monomeric (Klose and Bird, 2004).

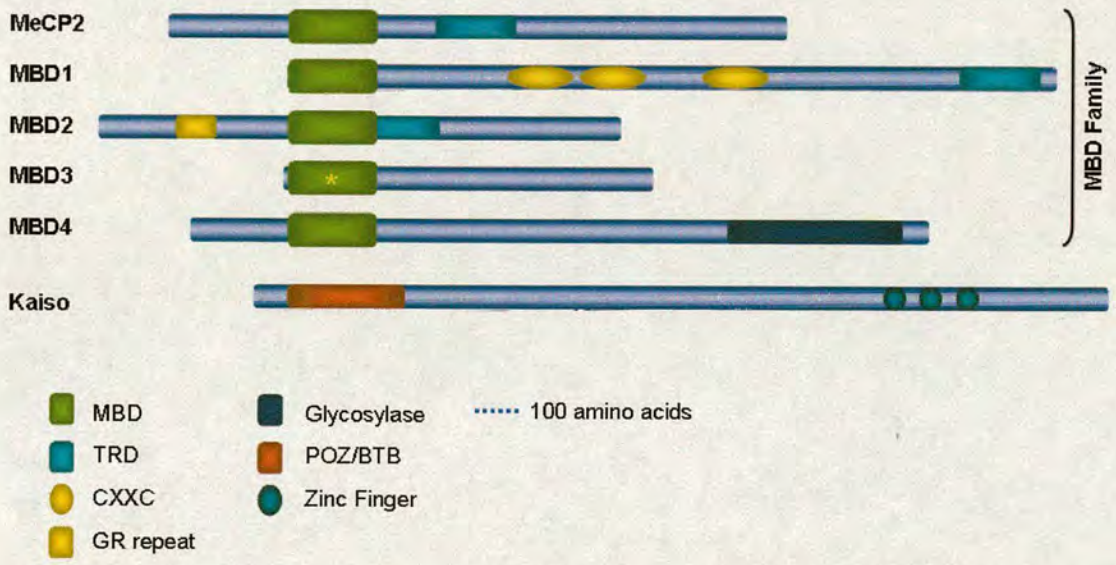


Figure 1.2-7. Mammalian Methyl Binding Proteins
MeCP2 binds to methylated DNA via its MBD which is characteristic of the MBD family proteins (bracketed). This mediates transcriptional repression through a C terminal TRD. MBD1 is similar but can target unmethylated DNA sequences through interaction with one of the CXXC motifs (CXXC). MBD2 also binds DNA via its MBD and mediates repression via interactions with a TRD. MBD2 contains unstructured Glycine / Arginine repeats (RG). MBD3 has an N terminal MBD domain although a Tyrosine to Phenylalanine transition (asterisked) prevents specific DNA binding activity. MBD4 is a mismatch repair protein, which binds CpG deamination products and excises the mismatched base pair via the C terminal glycosylase domain (Glycosylase). Kaiso is unique amongst the MBPs as it binds DNA through a zinc finger motif (zinc finger) rather than an MBD. Kaiso, mediates transcriptional repression through its POZ/BTB domain (POZ/BTB).

MeCP2 levels are particularly high in neuronal cells, consistent with the observation that mutations in MeCP2 give rise to the human neurological disorder; Rett Syndrome (Amir et al., 1999; Nan and Bird, 2001). A mouse deficient for MeCP2 provides a convincing model for the disease, presenting postpartum onset of both learning and cognitive impairment consistent with the disease phenotype in humans (Guy et al., 2001). Interestingly, a recent

study indicated that restoration of a functional protein in a transgenic mouse could rescue the majority of the diseased phenotype (Guy et al., 2007). This suggests that the *in vivo* localisation of MeCP2 is hard wired into the chromatin allowing MeCP2 to decipher DNA methylation and restore wild type gene expression.

Further to its role as a mediator of transcriptional repression, MeCP2 has been implicated in functions such as RNA splicing and chromatin organization (Georgel et al., 2003; Nikitina et al., 2007; Young et al., 2005). The fact that endogenous gene expression levels across the whole genome are only modestly affected by MeCP2 deficiency suggests that it may possess additional, as yet, uncharacterised cellular functions (Tudor et al., 2002). A recent study investigating global MeCP2 occupancy indicated preferential localisation to expressed gene promoters in a human neuronal cell line (Yasui et al., 2007). This finding is somewhat at odds with the historical role of MeCP2, and will require further investigation to determine its biological significance.

MBD1

MBD1 is a transcriptional repressor comprising an MBD, TRD and two or three conserved cysteine rich CXXC domains (discussed in chapter 3;(Cross et al., 1997b; Hendrich and Bird, 1998)). MBD1 colocalises to heterochromatic genomic foci but the localisation is not perturbed in DNMT deficient cells (Hendrich and Bird, 1998; Jorgensen et al., 2004). This methylation independent affinity is facilitated by the CXXC-3 domain which binds nonmethylated sequences *in vitro* and *in vivo* (Jorgensen et al., 2004). Accordingly, transcriptional repression can be achieved from both methylated and nonmethylated DNA templates (Fujita et al., 1999; Jorgensen et al., 2004). Human MBD1 has been shown to associate with the histone methyltransferase SETDB1 and transiently with chromatin assembly factor (CAF1) during S phase. Ablation of MBD1 in HeLa cells, leads to the hyperacetylation of certain target gene promoters (Sarraf and Stancheva, 2004). These findings indicate the existence of a coordinated propagation mechanism between DNA methylation and histone modification. This would suggest that newly deposited nucleosomes are methylated on H3K9 in response to DNA methylation via an indirect PCNA interaction (Sarraf and Stancheva, 2004). The formation of this complex may be regulated by the posttranslational modification of MBD1 by the small ubiquitin like modifier (SUMO; (Lyst et al., 2006)).

MBD2 and 3

MBD2 and MBD3 are the most closely related amongst the MBD proteins showing an overall amino acid sequence conservation of approximately 70% (Hendrich and Bird, 1998). MBD2 can bind methylated DNA *in vitro* and *in vivo* but a tyrosine to phenylalanine substitution has ablated specific DNA binding for MBD3 (Fraga et al., 2003; Hendrich and Bird, 1998). Both MBDs associate with the Nucleosome Remodeling and Histone Deacetylase (NuRD) corepressor complex (Bowen et al., 2004; Muchardt and Yaniv, 1999; Zhang et al., 1999). Initial results indicated that the two family members resided within the same repressive complex (Zhang et al., 1999). Genetic analysis has indicated that the MBD2 and MBD3 are not functionally redundant which has subsequently led to the biochemical purification of MBD2 and MBD3 specific NuRD complexes (Hendrich et al., 2001; Le Guezennec et al., 2006). This discovery may indicate that these MBDs play a role in distinct transcriptional repression in response to different physiological conditions or cellular processes.

MBD2 deficient mice are viable, fertile and present only a mild maternal nurturing defect (Hendrich et al., 2001). Molecular analysis has indicated that MBD2 deficiency results in slight de-repression of the *XIST* transcript and ectopic expression of a panel of genes in the colon (Barr et al., 2007; Berger et al., 2007). Conversely, MBD3 deficiency is embryonic lethal prior to implantation, suggesting that MBD3 is essential for completion of the developmental program (Hendrich et al., 2001). This phenotype appears to be mediated by an inability to repress the expression of specific pluripotency markers in the absence of a functional MBD3/NuRD complex (Kaji et al., 2006; Kaji et al., 2007).

MBD4

Whilst DNA methylation is essential for mammalian development, it represents a significant mutational burden (Li et al., 1992; Okano et al., 1999). 5MeC is hyper-mutable as it is prone to spontaneous deamination to thymine (Bird, 1980). Indeed approximately 1/3rd of human genetic disease have been attributed to such deamination events (Cooper and Krawczak, 1990). MBD4 can bind to both fully and hemimethylated CpG sites but has elevated affinity for T/G mismatches which are the product of ^{me}CpG deamination (Hendrich et al., 1999). Furthermore, MBD4 possesses a carboxy terminal glycosylase domain which can excise thymine and uracil when paired with a guanine base *in vitro* (Hendrich et al., 1999). Consistent with this observation, MBD4 deficiency results in a three fold over representation of CG to TA transition mutations (Millar et al., 2002). Deficient mice show an elevated

tumour burden and increased mortality when crossed onto an $APC^{min/+}$ background (Millar et al., 2002). Mismatch mutations can disrupt genomic integrity characteristic of cancers which present a microsatellite instability phenotype. Accordingly, human cancers of this type have a significant association with mutations in the endogenous *MBD4* gene, although recent evidence suggests that this may be consequential rather than causative (Bader et al., 2000; Pinto et al., 2008; Riccio et al., 1999). This is strong evidence that MBD4 functions as a mismatch repair protein *in vivo* and has evolved to combat the intrinsic mutability imposed by a heavily methylated genome.

Further to this well characterised function, more recent evidence has suggested that like its kin, MBD4 can also act as a transcriptional repressor (Kondo et al., 2005). This study indicated that transcriptional repression was mediated by interaction with Sin3A and HDAC1 (Kondo et al., 2005). Further investigation will be required to determine if this observation translates into a genuine regulatory function at endogenous loci.

Kaiso

Kaiso is an MBP of the BTB/POZ transcription factor family which bears no sequence similarities with the MBD proteins (Prokhortchouk et al., 2001). DNA binding is achieved through the recognition of two symmetrically methylated CpGs or a nonmethylated consensus site through interaction with a conserved zinc finger domain (Fig. 1.2-7) (Daniel et al., 2002; Prokhortchouk et al., 2001)). Transfection assays identified Kaiso as a methyl specific repressor in *Xenopus* and ablation of the protein facilitated precocious transcription of a panel of developmental genes (Ruzov et al., 2004). It is unclear whether the affinity for the nonmethylated motif in mammalian cells can also direct sequence specific, methylation-independent transcriptional repression (Daniel et al., 2002). Kaiso associates with the HDAC dependent NCoR corepressor complex through interaction between the N terminal POZ domain of Kaiso and the RD1 domain of hNCoR. A functional kaiso/NCor complex is required for repression of endogenous *MTA2* expression (Yoon et al., 2003). Disruption of the complex by siRNA depletion of NCoR or Kaiso, leads to derepression of the *MAT2* and concomitant histone hyperacetylation and the loss of H3K9 methylation (Yoon et al., 2003).

A degree of redundancy between the MBPs is suggested by their functional and mechanistic similarities. This proposal is supported by the mild phenotype presented by MBD2 deficiency and the weak genetic interaction observed between MBD2 and 3 (Hendrich et al., 2001). However, the observation that MBPs have different in vitro sequence preferences

suggests that they are not completely functionally equivalent (Table 1.2-1). Interestingly, the single member of the family lacking methyl CpG binding (MBD3) is embryonic lethal which suggests its function extends beyond sensing the methylation state of DNA alone. The various MBPs seem to have evolved to capitalize, decipher and protect the DNA methylation mechanism. For an insight into the evolution of DNA methylation and the MBD protein see review by Colot and Rossignol and references therein (Colot and Rossignol, 1999).

There is evidence indicating that MeCP2, MBD2-4 and Kaiso are HDAC dependent transcriptional repressors. Furthermore the majority of MBPs have been associated with H3K9 methyltransferase activity, a histone mark known to associate with inactive chromatin (Bernstein et al., 2007). These findings indicate that modulation of chromatin structure is a primary mechanism employed by MBPs, but that this does not exclude a secondary role in the steric hindrance of transcription factor binding (Fig. 1.2-2).

Table 1.2-1. Sequence binding affinity for Methyl-Binding Proteins

MBP	Binding site	Methylation Status[a]	Binding Domain	Reference
MeCP2	CpG+A/T[≥4]	M	MBD	(Klose et al., 2005; Nan et al., 1993)
MBD1	CpG	M/N	MBD/CXXC	(Hendrich and Bird, 2000; Jorgensen et al., 2004)
MBD2	CpG	M	MBD	(Hendrich and Bird, 1998; Klose et al., 2005)
MBD3	NA	NA	NA	(Fraga et al., 2003; Hendrich and Bird, 1998)
MBD4	C/U/TpG	M/H/M	MBD	(Hendrich and Bird, 1998; Millar et al., 2002)
Kaiso	CpGpCpG or TCCTGCNA	M/N	Zinc Finger	(Daniel et al., 2002; Prokhortchouk et al., 2001)

[a]Methylation status of binding sequence are indicated as either methylated (M), nonmethylated (N), hemimethylated (H) or non applicable (NA).

1.2.8 Mammalian X-inactivation – An Epigenetic Paradigm

One of the best characterised examples of the functional importance of DNA methylation is mammalian X-inactivation. In mammals, sex determination is mediated by a pair of heteromorphic sex chromosomes. X inactivation is the process whereby females, bearing two X chromosomes, compensate for the genetic imbalance between the sexes. To equalize X-linked gene expression levels between males and females a single X chromosome is inactivated during early development (Heard et al., 1997). Interestingly, during the first five cell divisions of female embryogenesis, the paternal X chromosome (Xp) is inactivated. Subsequently, features of early imprinted X-inactivation are lost in the inner cell mass allowing random X-inactivation to proceed in the epiblast (Okamoto et al., 2004). DNA

methylation plays a central role in X-inactivation as disruption of *DNMT1* leads to severe perturbations in X inactivation inducing ectopic silencing of the active X chromosome (Xa; (Norris et al., 1994; Panning and Jaenisch, 1996)). Furthermore, transient depletion of DNA methylation, following 5AzaC treatment can reactivate the PGK-1 gene located on the Xi in hybrid mammalian cells (Hansen and Gartler, 1990).

The molecular process of random X inactivation, proceeds through a reversible initiation stage and then culminates in the irreversible repression of a single X chromosome (Wutz and Jaenisch, 2000). The *Xic* (X-inactivation centre) encodes *Xist,* a large ncRNA, which is essential for the initiation of X-inactivation (Penny et al., 1996). Xi specific expression of *Xist* follows inactivation choice, and initiates chromosome-wide repression *in cis* (Brown et al., 1991; Clemson et al., 1996; Wutz and Jaenisch, 2000). The restricted pattern of *Xist* expression is mediated by the *Tsix* antisense ncRNA, which is expressed from the Xa and is required for *Xist* repression (Luikenhuis et al., 2001; Sado et al., 2002). Furthermore, *Xist* is regulated, at least in part, by differential DNA methylation of its promoter region which is hypomethylated and hypermethylated on The Xi and Xa respectively in somatic cells (Norris et al., 1994). In support of this scenario, *DNMT1* deficiency was shown to reactivate *Xist* expression form the Xa, and elicit ectopic X inactivation (Norris et al., 1994; Panning and Jaenisch, 1996). *Xist* RNA physically coats the Xi, which facilitates transcriptional repression which occurs very shortly after *Xist* transcription is initiated (Clemson et al., 1996; Keohane et al., 1996). Interestingly, *Xist* stabilization is important for this process with the half life of the RNA transcript being significantly higher in somatic relative to ES cells (Sheardown et al., 1997). Interestingly, recent evidence suggests that early paternal specific X inactivation corresponds to specific gamete-derived DNA methylation patterns within the *Xic*, analogous to autosomal imprinting (Boumil et al., 2006). Unlike imprinting however, these methyl-marks are erased in the inner cell mass to allow subsequent random X-inactivation.

Following *Xist* mediated initiation events; subsequent inactivation involves the formation of facultative heterochromatin which occurs in response to *Xist* localisation. Artificial differentiation experiments provided much of the information pertaining to temporal order of these events (Brockdorff, 2002). Following the accumulation of *Xist*, the Xi becomes methylated at H3K9 and hypoacetylated on H3K9 and H3K4 (Heard et al., 2001; Mermoud et al., 2002). This corresponds with a shift to late replication of the Xi during S-phase (Keohane et al., 1996; Priest et al., 1967). Subsequent to asynchronous replication, histone

H4 tails become hypoacetylated which may contribute to the stabilization of inactivation required for the transition between initiation and maintenance (Heard et al., 2001; Keohane et al., 1996). The H2A1.2 histone variant is deposited on the Xi, although interestingly this was found to be reversible if *Xist* was depleted in MEFs (Csankovszki et al., 1999). Consistent with this result, depletion of *Xist* in ES cells prior to 72 hours was shown to reverse X-inactivation with the concomitant loss of early chromatin modifications. Once ES cell differentiation extends beyond this point, inactivation becomes stable and independent of *Xist* expression (Wutz and Jaenisch, 2000). However there is evidence that *Xist* can contribute, in part, to the maintenance of X inactivation (Csankovszki et al., 2001). Synthetic depletion of *Xist* in MEFs results in partial relief of transcriptional silencing on the Xi. The same study indicated that efficient maintenance of the inactive state requires *Xist*, DNA methylation and hypoacetylation of histone tails for efficient transcriptional repression (Csankovszki et al., 2001).

Further to the modifications already discussed, polycomb repressive complexes play a central role in X inactivation. Components of both PRC1 and PRC2 complexes associate with the Xi very shortly after *Xist* expression (de Napoles et al., 2004; Plath et al., 2003; Silva et al., 2003). Concurrently, histone H2AK9 is ubiquitinated and histone H3K27 is methylated (de Napoles et al., 2004; Plath et al., 2003; Schoeftner et al., 2006). The PRC2 complex, and H3K27me2 and me3 \were found to be indispensable for Xi gene silencing. Moreover, H2A ubiquitination was found to be independent of PRC2, despite the fact that H3K27 methylation is known to recruit PRC1 (Schoeftner et al., 2006). Moreover PRC1 components could be recruited to DNA in an *Xist* dependent and PRC2 independent manner (Schoeftner et al., 2006). Alternatively, disruption of DNA methylation led to mislocalisation of the PRC1 component BMI1, but did not affect H3K27methylation (Hernandez-Munoz et al., 2005). These data suggest that H3K27methylation, DNA methylation and *Xist* may have an overlapping role in PRC1 recruitment.

Maintenance of transcriptional repression corresponds with promoter-CGI methylation and hypoacetylation on the Xi (Gilbert and Sharp, 1999; Norris et al., 1991; Tribioli et al., 1992; Weber et al., 2005; Wolf et al., 1984; Yen et al., 1984). Despite the fact that CGI methylation correlates with transcriptional repression, data suggests that it is required for stable maintenance rather than initiation of repression. This is supported by the observation that Xi genes are already repressed prior to the acquisition of DNA methylation at these sites (Keohane et al., 1996). Interestingly, analysis of allelic methylation suggests that intergenic

regions are relatively hypomethylated on the Xi, although the biological relevance of this is unclear (Gilbert and Sharp, 1999; Weber et al., 2007; Yen et al., 1984). One hypothesis is that gene-body methylation associated with genes of the Xa, may serve to prevent spurious transcription from internal cryptic promoters or transposable elements. Such internal expression has been proposed to interfere were correct transcription as previously discussed (Suzuki and Bird, 2008). However analysis of Xa specific gene body methylation indicated no correlation with genes bearing internal repetitive elements (Hellman and Chess, 2007). Alternatively, this may simply represent the global hypomethylation of the inactive X-chromosome relative to its active counterpart (Viegas-Pequignot et al., 1988; Weber et al., 2005).

Gene repression on the Xi is not ubiquitous as many genes avoid repression and are biallelically expressed form both chromosomes (Brown and Greally, 2003; Carrel et al., 1996; Carrel and Willard, 2005). Active genes located on the Xi are generally localised to the region that is homologous with the Y chromosome. It is therefore possible that these represent genes which can opt out of inactivation due to equal dosage between the sexes (Brown et al., 1991). Alternatively, genes escaping X-inactivation have been shown to correlate with more recent acquisition to the sex chromosome, via autosomal translocation. This would suggest an alternative hypothesis, where selective pressure has not yet dictated the inactivation of these genes (Graves, 2006). The mechanism by which these genes remain transcriptionally active is unclear. A recent microarray study investigating the DNA methylation status of human gene promoters indicated that promoters of genes which escape inactivation were hypomethylated relative to their inactivated counterparts, consistent with previous reports (Carrel et al., 1996; Weber et al., 2007). This is particularly interesting, since it has been posited that these genes contribute to sexual dimorphic traits (Carrel and Willard, 2005). Furthermore, escaping genes are somewhat variable between individuals, which may consequently provide female specific phenotypic heterogeneity (Carrel and Willard, 2005). These findings suggest a possible role for DNA methylation in establishing or propagating these polymorphic traits.

Haploinsufficiency occurs when a diploid organism transcribes a gene from only one allele, such that expression is insufficient for correct gene function. Consequently, X inactivation represents a paradox, as it appears to impose this fate by creating X chromosome specific monosomy, which results in Turner Syndrome in humans (Zinn and Ross, 1998). However, there is recent evidence to suggest, that mammalian cells avoid this deleterious effect, by

specifically up-regulating the Xa, following zygote formation (Nguyen and Disteche, 2006). The mechanisms facilitating this process are unclear at present, but this represents an unanticipated aspect of mammalian dosage compensation.

1.3 CpG Islands

The mammalian genome is globally methylated at the majority of CpG sites with methylation being localised to gene bodies, endogenous repeats and transposable elements (Eckhardt et al., 2006; Ehrlich et al., 1982; Weber et al., 2005). However, this methylated landscape is punctuated by CpG rich non-methylated sequences, called CGIs (Bird et al., 1985; Cooper et al., 1983; Gardiner-Garden and Frommer, 1987). CGIs have an increased G+C content and are enriched for CpG dinucleotides, although this CpG density alone is insufficient to account for the elevated GC composition. Suppression of CpG, characteristic of the majority of the genome, is not observed in these regions due to the lack of DNA methylation and the consequent absence of deamination. The existence of CGIs resolves the apparent discrepancy whereby 15% of CpG sites localise to approximately 2% of the mammalian genome (Antequera and Bird, 1993; Ehrlich et al., 1982).

This human CGI fraction was first identified by digesting genomic DNA with the methyl-sensitive restriction endonuclease HpaII (Antequera and Bird, 1993; Cooper et al., 1983). This preparation preserved the majority of the genome as high molecular weight DNA, whilst generating a small proportion of highly restricted products termed HTFs (HpaII Tiny fragments; (Antequera and Bird, 1993; Cooper et al., 1983)). Based on the characterisation of these initial digestion products, CGIs were defined as sequences longer than 200bp which have a G+C composition of more than 50% and a CpG[o/e] in excess of 0.6 (Gardiner-Garden and Frommer, 1987; Larsen et al., 1992). These parameters have been refined based on data generated by the human genome project, however this definition is still widely used for sequence based CGI prediction (computational CGI prediction will be discussed later; (Hackenberg et al., 2006; Lander et al., 2001; Ponger and Mouchiroud, 2002; Takai and Jones, 2002; Waterston et al., 2002)).

CGIs are, on average, one kilobase in length and notably overlap the promoter regions of approximately 60-70% of all human genes[vii] (Bird et al., 1985; Lander et al., 2001; Larsen et

[vii] The actual number is likely to be closer to 60% as CGIs are used to predict gene promoters. Therefore, annotated genes will be artificially biased towards the promoter CGI category Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG

al., 1992; Saxonov et al., 2006; Weber et al., 2007). Preliminary chromatographic purification of these sequences was driven by their utility in mapping cDNA (cloned DNA) sequences to their 5' regulatory elements (Cross et al., 1994). Specifically, CGIs have been shown to colocalise with the promoters of all housekeeping genes and approximately 40% of those displaying a tissue restricted expression profile (Larsen et al., 1992). The more ubiquitously expressed CGI promoters appear to define a class of transcription start sites which can initiate from multiple positions within the promoter. The more tissue restricted class on the other hand is generally associated with a single well defined initiation site and a TATA consensus (reviewed in (Sandelin et al., 2007)). Promoter association accounts for the uneven distribution of CGIs in the genome, showing preferential localisation to gene rich regions (Craig and Bickmore, 1994; Lander et al., 2001). A combination of molecular and computational mapping methodologies have determined that there are between twenty and thirty thousand CGIs in the haploid human genome (Antequera and Bird, 1993; Ewing and Green, 2000; Lander et al., 2001; Takai and Jones, 2002).

Consistent with promoter association, CGIs are characterised by a transcriptionally permissive chromatin state (Bird, 1987; Larsen et al., 1992; Tazi and Bird, 1990). These findings suggest that CGIs may provide a means to distinguish gene promoter regions from the large proportion of transcriptionally irrelevant intergenic chromatin. This would suggest that CGIs facilitate the correct association of ubiquitous transcription factors to their cognate sites within regulatory sequences. Support for this idea was provided by an early study investigating the distribution of transcription factor binding sites in a small panel of human genes (Prestridge and Burks, 1993). Whilst binding sites were slightly enriched in promoter proximal sequences, they were also highly abundant throughout the genome (approximately 16 sites per 100bp). This study concluded that the presence of binding sites alone was insufficient to identify promoters, which supports the idea that CGIs may serve as TF 'landing lights' in the darkness of the nucleus (Bird, 1995; Prestridge and Burks, 1993).

CGIs do not always localize to annotated transcriptional start sites of protein coding genes. However, is interesting to note that previous detailed investigation of intergenic, 'orphan CGIs', has subsequently led to the identification of previously unanticipated promoters (Gardiner-Garden and Frommer, 1994; Kleinjan et al., 2004; Macleod et al., 1998). This suggests that all CGIs may represent sites of transcriptional initiation which have thus far

---

dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A *103*, 1412-1417..

remained unidentified. Indeed it is possible that certain alternative start sites are utilised in a highly tissue restricted fashion, and as such have escaped annotation. Indeed several genes which are transcribed from intragenic CGIs have been found to be expressed during highly specific developmental stages (Kleinjan et al., 2004; Macleod et al., 1998).

1.3.1 The Origin and Maintenance of CpG Islands

The characteristic clustering of CpG sites is a consequence of CGI immunity against *de novo* methylation during the earliest stages of mammalian development. However the mechanism by which CGIs remain hypomethylated during the period of global *de novo* methylation remains unclear. A simple suggestion would be that they are intrinsically refractive to *de novo* methylation by preventing the action of the DNMTs. This seems unlikely however, as CGIs contain a substantially elevated density of CpG sites, the preferred substrate of the DNMT enzymes (Ramsahoye et al., 2000). Moreover, CGIs located on the female inactive X chromosome and those of certain cultured mammalian cells readily acquire DNA methylation (Antequera et al., 1990; Weber et al., 2005).

A plausible alternative is that bound transcription factors sterically preclude the activity of DNMTs at CGI sequences (Cuadrado et al., 2001). Evidence for such a mechanism is supported by mouse transgenic experiments in which ablation of binding sites for the ubiquitous transcription factor Sp1 was shown to facilitate *de novo* methylation of the *APRT* promoter CGI (Brandeis et al., 1994; Macleod et al., 1994). Consistently, α-globin which is transcribed in the embryo contains a promoter CGI whilst the related, transcriptionally silent, β-globin gene does not (Daniels et al., 1997). Moreover, analysis of a panel of genes expressed during mouse embryogenesis found that 93% are associated with a 5' CGI (Ponger et al., 2001). A global characterisation of histone modifications in mouse ES cells identified islands of H3K4me3 at almost all CGIs (Mikkelsen et al., 2007). These data suggest that CGIs are footprints of the basal transcription machinery localised during embryogenic *de novo* methylation. However, this model does not account for the observation that CGIs are intrinsically sensitive to nuclease digestion and therefore more accessible than the majority of the genome (Antequera et al., 1989; Tazi and Bird, 1990). Moreover, the majority of CGIs remain hypomethylated in terminally differentiated cells irrespective of gene association and transcriptional activity (Eckhardt et al., 2006; Larsen et al., 1992; Weber et al., 2007).

A third possibility is that CGIs are targeted by a DNA demethylation mechanism, which specifically removes the methyl moiety from the cytosine base. Various protein factors,

including CGBP (CpG Binding Protein) possess a CXXC domain, which can specifically bind to non-methylated CpG sites (Carlone et al., 2002; Voo et al., 2000). This protein has been shown to associate with the MLL complex, which mediates the formation of transcriptionally permissive chromatin via histone modifying activities (Ansari et al., 2008). It is possible, that an equivalent recruitment mechanism could target a demethylation activity to CGIs. However, no such demethylase activity has thus far been identified in somatic tissues.

Recent evidence suggests a rather speculative alternative involving the methyltransferase like factor DNMT3L. As previously discussed, this protein can associate with, and facilitate the action of the *de novo* methyltransferases (Gowher et al., 2005; Hata et al., 2002; Jia et al., 2007; Suetake et al., 2004). However, this protein cannot bind to chromatin in which the Histone H3 tails are tri-methylated at the lysine 4 position (Ooi et al., 2007). A recent investigation indicated that the majority of protein coding gene promoters were occupied by RNA polymerase II and trimethylated at H3K4 even in the absence of transcriptional elongation (Guenther et al., 2007). The presence of this active mark at CGI-promoters may prevent *de novo* methylation via repulsion of DNMT3L. Moreover, the persistence of these modifications in somatic tissues where transcription levels are below detection may still prevent accumulation of methylation through the same mechanism. Indeed, one study indicated that tissue specific genes may be expressed at very low levels, but are nonetheless active (Cuadrado et al., 2001). This model is speculative at this point and requires experimental scrutiny to determine its validity.

High density clustering of CpG sites alone does not account for the atypical G+C composition of mammalian CGIs. Moreover, it is unlikely that the basal transcription machinery alone could generate a 1 kilobase footprint characteristic of CGI sequences. Interestingly, several vertebrate CGIs are found to associate with origins of DNA replication (Delgado et al., 1998; Phi-van and Stratling, 1999; Rein et al., 1997). This led to the proposal that both elevated G+C composition and lack of DNA methylation could arise through the association of CGIs with altered DNA metabolism at these sequences (Antequera and Bird, 1999). The unique looping of DNA during replication initiation in combination with the replication apparatus may provide an alternative means by which DNA methylation could be precluded from these sites (Antequera and Bird, 1999).

Several of these models are likely to be involved in the establishment of CGIs and the hypomethylation which persists during subsequent differentiation. High throughput analysis of chromatin modifications; transcriptional activity; transcription factor binding and global methylation analysis will likely provide further insight into the origin of CGIs in the future.

1.3.2 Aberrant CpG Island Methylation

Knudson's 'two hit hypothesis' dictates that incremental mutations are required for the conversion of 'normal' cells into highly proliferative cells characteristic of cancer. In this he stated that two or more mutational events were required to disrupt cellular regulation and consequently facilitate cancer formation (Knudson, 1971). At this stage cancer was considered as a genetic disorder, resulting from mutations in the underlying DNA sequence itself. However, more recent data suggests that epigenetic abnormalities also play an important role in the etiology of neoplasia (reviewed in (Jones, 2002; Jones and Baylin, 2007; Ting et al., 2006)).

Cancer cells and transformed cell lines are often characterised by a global paucity of DNA methylation (see for example (Rodriguez et al., 2006; Weber et al., 2005)). This has been implicated in genome instability, and subsequent genetic abnormalities such as chromosomal rearrangements and aneuploidy (Fukasawa, 2005; Rodriguez et al., 2006). Alternatively, there is extensive evidence indicating that unscheduled *de novo* methylation of CGIs associated with tumour suppressor promoters is a frequent hallmark of many cancer cells (Huang et al., 1999; Weber et al., 2005; Yan et al., 2002; Yan et al., 2000). Consequently, epigenetic silencing (epimutation) of tumour suppressor genes has been suggested as a primary event leading to unchecked proliferation associated with metastasizing cells. Interestingly, a similar phenomenon has been reported in permanent cultured cell lines, where CGIs associated with non-essential gene promoters are frequently *de novo* methylated (Antequera et al., 1990).

Is cancer specific CGI methylation a random stochastic event or a more targeted phenomenon directed towards particular DNA sequences? Evidential support for the former comes from the observation that several DNA sequences aberrantly acquire DNA methylation during the process of aging in mammalian cells (Issa et al., 1996; Kwabi-Addo et al., 2007; Oakes et al., 2003; Toyota et al., 1999). A study of epigenetic variation in monozygotic twins, determined that global and regional methylation levels varied more extensively within older (>28) twin pairs (Fraga et al., 2005). This phenomenon could result

in an accumulation of ^{me}CpG at promoter CGIs inadvertently invoking gene silencing of tumour suppressors and provide a selective growth advantage to these cells. However, these studies are based on a small numbers of candidate loci or are relatively low resolution with respect to individual CpG sites. Consequently changes in DNA methylation levels as a function of age, has proven to be contentious. A large scale bisulfite sequencing analysis of 512 human CGIs failed to identify significant methylation differences between two panels of tissues of disparate age (mean ages of 26 and 68; (Eckhardt et al., 2006)). Moreover, there is evidence to suggest that cancer specific methylation is distinct from that associated with human aging (Toyota et al., 1999). Analysis of a panel of colorectal cancers indicated that CGI methylation occurred in both metastatic cells and adjacent parenchyma in an age dependent fashion. However, certain CGIs were found to be methylated specifically in a subset of cancer cells and were distinct from those found in normal parenchyma. This phenomenon was termed the CpG Island Methylator Phenotype (CIMP; (Toyota et al., 1999)). This finding alone does not however, rule out the possibility of random CGI methylation, and clonal selection and expansion. More recently several studies have identified a link between chromatin modifications and CGIs which acquire aberrant methylation in cancer (Keshet et al., 2006; Ohm et al., 2007; Schlesinger et al., 2007). H3K27me3 is a repressive histone modification associated with the promoters of many developmental genes during the early stages of embryogenesis. Interestingly, islands which were found to be aberrantly methylated in cancer cells and cell lines were found to be tightly associated with genes marked by this modification in ES cells (Keshet et al., 2006; Ohm et al., 2007; Schlesinger et al., 2007). These studies concluded that DNA methylation may serve to lock in a pseudo pluripotent stem cell like state, which facilitates cellular proliferation. These findings suggest an "instructive" mechanism whereby this chromatin modification directs cancer specific methylation; although it is unclear whether this is the only mechanism giving rise to aberrant CGI methylation (Keshet et al., 2006; Ohm and Baylin, 2007; Ohm et al., 2007; Schlesinger et al., 2007).

1.3.3 'Normal' CpG Island Methylation

Despite this wealth of data pertaining to the methylation status of CGIs in cancers, relatively little attention has been paid to the equivalent phenomenon in 'normal' cells. Indeed, a small but significant proportion of CGIs does acquire methylation during human development, and is implicated in functionally important processes such as X-inactivation (previously discussed) and parental imprinting (Reik, 2007). CGI methylation at imprinted loci is generally chromosome specific and has a functional role in the regulation of monoallelic

expression (Li et al., 1993; Sleutels et al., 2002; Wutz et al., 1997)}. A CGI located at the *IGF2R/H19* locus serves as an insulator element, which can either promote or repress the expression of *IGF2R* in a methyl-dependent manner. Moreover, the non coding RNA *Air* is paternally expressed from the differentially methylated intragenic CGI of murine *Igf2r*, and represses transcription of the maternally expressed gene cluster (Sleutels et al., 2002; Wutz et al., 1997). The importance of correct methylation at this CGI has been illustrated by the complete repression of certain *IGF2R* isoforms, as a consequence of biallelic methylation during aging and carcinogenesis in human cells (Issa et al., 1996). A third class of methylated islands has been characterised, and these associate with the promoter regions of germ line specific genes which are silenced in somatic tissues (De Smet et al., 1999; Shen et al., 2007; Weber et al., 2007). Genes of the *MAGE* (Melanoma Antigen Encoding Genes) family are silenced during embryogenesis primarily through the methylation of CpG rich promoter sequences (De Smet et al., 1999). Accordingly, promoter demethylation correlates with ectopic expression in various cancer cells (Honda et al., 2004).

Several studies have identified additional methylated CGIs (Eckhardt et al., 2006; Kitamura et al., 2007; Strichman-Almashanu et al., 2002; Weber et al., 2007; Yamada et al., 2004). A digestion based cloning strategy identified 43 hypermethylated CGIs in human somatic tissues (Strichman-Almashanu et al., 2002). Interestingly, detailed analysis determined that approximately half of these sequences were hypermethylated in germ-line derived cells from both parents, indicating that these sequences do not represent imprinted loci (Strichman-Almashanu et al., 2002). This finding was unexpected as methylation during this developmental stage is likely to result in heritable depletion of CpG sites through spontaneous mutation of methyl-cytosine. Methylation analysis of all predicted[viii] CGIs (149) on the q arm of human Chromosome 21 determined that 22% were heavily methylated in peripheral blood DNA (Yamada et al., 2004).

More recently, high throughput analysis has provided a more comprehensive view of the human CGI methylome (Eckhardt et al., 2006; Weber et al., 2007). Weber and coworkers screened the majority of human gene promoters by probing oligonucleotide arrays with DNA prepared using the MeDIP technique (Weber et al., 2005; Weber et al., 2007). Interestingly,

[viii] Prediction was based on the formal CGI criteria proposed by Gardiner-Garden and Frommer, with an increased stringency for CGI length (>400bp) and repeat masking Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J Mol Biol *196*, 261-282, Yamada, Y., Watanabe, H., Miura, F., Soejima, H., Uchiyama, M., Iwasaka, T., Mukai, T., Sakaki, Y., and Ito, T. (2004). A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. Genome Res *14*, 247-266..

relatively few CGI[ix] promoters were found to be methylated (~3%) in somatic tissues and primary cell lines (Weber et al., 2007). These finding were consistent with those of an independent study, which identified 4% of CGI promoters as hypermethylated in whole peripheral blood leukocytes (Shen et al., 2007). Alternatively, promoters with relatively reduced CpG content were frequently found to be methylated in these cells (Weber et al., 2007). This finding is consistent with the observation that a methylated fraction purified from human whole blood was found to be enriched for sequences with a CpG density intermediate between CGIs and bulk genomic DNA (Brock et al., 1999). High resolution bisulfite sequencing analysis was carried out for 2,524 regions of human chromosomes 6, 20 and 22 in a panel of 12 tissues (Eckhardt et al., 2006). This study identified 9.2% of predicted CGIs as methylated at more than 80% of CpG sites in one or more somatic tissues (Eckhardt et al., 2006). The discrepancy between the observed numbers from these two studies likely arises due to two key experimental factors. 1). Eckhardt and colleagues partitioned methylation into 3 categories including those sequences presenting intermediate levels of methylation (>20 and <80% ^{me}CpG). 2). The bisulfite analysis assayed CGIs irrespective of promoter association, and therefore may reflect an altered tendency towards methylation of these sequences (Eckhardt et al., 2006; Weber et al., 2005). Consistently, only 2.1% of promoter associated CGIs were identified as hypermethylated (>80%) by bisulfite sequence analysis (Eckhardt et al., 2006). Furthermore, both studies confirmed that many CGIs in sperm are hypomethylated with respect to those in somatic tissues (Eckhardt et al., 2006; Weber et al., 2005).

A small proportion of CGIs presented tissue specific differential methylation levels (Eckhardt et al., 2006). This finding is particularly interesting as it is conceivable that differential CGI methylation could function in tissue specific gene regulation. Consistently, candidate analysis of the CpG rich promoter of the human gene *MASPIN* identified tissue specific methylation. Promoter methylation levels were shown to correlate with transcriptional inactivity (Futscher et al., 2002). RLGS (Restriction Landmark Genomic Scanning) analysis of DNA methylation in five somatic tissues and mature sperm in mice indicated that as many as 5% of CGIs associated with tissue specific methylation profiles. Analysis of 14 candidate loci determined that the majority were specifically hypomethylated

[ix] CGIs in this study are referred to as HCPs (High CpG Promoters) whereas other ICPs (Intermediate CpG Promoters) have a sequence composition intermediate to that of CGIs and bulk genomic DNA Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M., and Schubeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet *39*, 457-466..

in sperm, although one CGI associated with the promoter of *Gata2* was specifically methylated in liver consistent with low level transcription in this tissue (Song et al., 2005).

Interestingly a recent study identified a panel of methylated promoters in murine ES cells, of which, approximately 3% were associated with CGIs (Fouse et al., 2008). These included members of the Reproductive homeobox (rHOX) class of transcription factors. The 0.7Mb rHOX cluster has been shown to associate with lineage specific methylation patterns which dictate the expression of the respective genes in the embryonic lineages (Oda et al., 2006). Analysis of DNMT mutants confirmed that these genes are primarily silenced by DNA methylation (Fouse et al., 2008; Oda et al., 2006). Minimal overlap between genes repressed by DNA promoter methylation and those targeted by PcG and Nanog / Oct4 suggests that there are multiple complementary regulatory mechanisms which maintain correct expression in the preimplantation embryo (Fouse et al., 2008).

These recent investigations have provided tantalising insights into distribution of methylation across this functionally important fraction of the genome. However, there are many remaining questions. Do tissue specific methylation patterns have a mechanistic role in hard wiring expression patterns in terminally differentiated cells? Does differential CGI methylation occur between equivalent healthy cells in different individuals? These questions must be addressed before we can begin to understand the role of this epigenetic phenomenon in transcriptional regulation and consequently the aberrant events associated with cancer.

1.4 PhD Objectives

The purpose of this PhD is to provide better insight into the distribution and methylation status of CpG islands in normal somatic tissues. This will be approached by two related chromatographic methodologies. In the first instance a novel purification technique will be developed to enrich for sequences containing clusters of nonmethylated CpG sites. This will be applied to the generation of a CGI set from whole human blood genomic DNA. Characterisation of this library will be used to determine the spatial distribution of CGIs relative to protein coding genes and with those identified by sequence based prediction algorithms. Subsequent analysis will focus on methylated CGIs in normal human tissues, as this remains a relatively poorly defined fraction of the genome. To this end, methylated CGIs will be purified from total genomic DNA using MAP chromatography (Brock et al., 1999; Cross et al., 1994). A microarray representing the entire CGI set will be probed with MAP fractions prepared from blood, brain, muscle and spleen DNA. Microarray results will be

validated using conventional techniques, and then used to assess genomic distribution, tissue variability and gene association of these sequences. It is hoped that these results will provide insight into previously unanticipated biological processes involving methylation of CGIs.

The majority of data presented in this thesis has been published and can be found in Appendix 1 (Illingworth et al., 2008).

Chapter 2: Materials and Methods

2.1 Materials and Reagents

All material and reagents were stored at room temeperature (r/t) unless otherwise stated. Standard Molecular Biological reagents will not be referred to here

2.1.1 DNA Manipulation

Orange G loading buffer (6×): 0.198% (w/v) orange G, 12% (w/v) Ficoll, 120mM EDTA pH8.0, 4.2% (w/v) SDS. Stored at –20°C (long-term) or r/t (short-term).

Proteinase K stock solution: 20mg/ml proteinase K, 100mM EDTA pH 7.5, 2% (w/v) SDS. Stored at –20°C.

DNA Sequencing Buffer (2.5x): 20mM Tris HCl (pH8) and 5mM MgCl2.

TAE electrophoresis buffer (1x): 40mM Tris-acetate, 1mM EDTA

TBE electrophoresis buffer (1x): 45mM Tris-borate, 1mM EDTA

Bisulfite modification solution (pH of 5.0): 3.8g sodium hydrogen sulfite ($NaHSO_3$) was dissolved in 5ml dH_2O and 1.5ml 2M NaOH (protected from light). 110mg hydroquinone was dissolved in 1ml dH_2O at 55°C for ten minutes. The sodium bisulphite and the hydroquinone solutions were then mixed. Bisulphite modification solution was prepared immediately prior to use.

TE buffer pH7.5: 10mM Tris HCl p7.5, 1mM EDTA

RCL Buffer (Red Cell Lysis): 150mM Ammonium Chloride (NH4Cl), 10mM Potassium Hydrogen Carbonate (KHCO3) and 0.01% (w/v) EDTA.

SL Buffer (Sperm Lysis): 6M Guanidinium Hydrochloride, 30mM Sodium Citrate (pH7), 0.5% (w/v) Sarkosyl, 0.2mg/ml proteinase K, 0.2mg/ml RNase A and 0.3M β-mercaptoethanol.

2.1.2 RNA Manipulation

10× MOPS buffer: 400mM MOPS, 100mM NaOAc, 10mM EDTA. pH adjusted to 7.0 with NaOH.

RNA loading buffer (3x): 60% (v/v) deionised formamamide, 8% formaldehyde, 0.5x MOPS buffer, 0.4% (w/v) bromophenol blue and 1mg/ml ethidium bromide. Stored at –20°C.

2.1.3 Protein Manipulation

Sodium Phosphate Buffer [pH8]: 6.8ml 1M Sodium dihydrogen phosphate (NaH2PO4) and 93.2 ml 1M di Sodium hydrogen phosphate (Na2HPO4).

PBS (1x): 140mM NaCl, 3mM KCl, 2mM KH2PO4, 10mM Na2HPO4

Chelating Sepharose wash buffer [pH4]: 0.02 M NaOAc and 1M NaCl.

N10 buffer: 50mM sodium phosphate buffer pH8.0, 300mM NaCl, 10% (v/v) glycerol, 10mM Imidazole, 15mM β-mercaptoethanol and 0.5mM PMSF. β-mercaptoethanol and PMSF were added immediately prior to use. Stored at 4°C

N20 buffer: As per N10 buffer but with 20mM Imidazole.

N250 buffer: As per N10 buffer but with 250mM Imidazole.

CEA buffer (Cation Exchange A): 20mM HEPES (pH7.9), 0.1% (v/v) Triton X-100, 10% (v/v) Glycerol, 0.5mM PMSF and 0.5mM β-mercaptoethanol. β-mercaptoethanol and PMSF were added immediately prior to use. Stored at 4°C.

CEB buffer (Cation Exchange B): As per CEA buffer with the addition of 1M NaCl.

Coomassie Brilliant Blue R-250 staining solution: 50% (v/v) methanol, 10% (v/v) glacial acetic acid, 0.1% (w/v) Coomassie Brilliant Blue R-250. Filtered through a Whatman number 1 filter.

Coomassie destain solution: 50% (v/v) methanol and 10% (v/v) glacial acetic acid.

CNBr coupling buffer (pH8.3): 0.1M NaHCO3 and 0.5M NaCl.

SDS PAGE loading buffer (2×): 100mM Tris HCl pH6.8, 4% (w/v) SDS, 20% (v/v) glycerol, 200mM DTT, 0.2% (w/v) bromophenol blue. Stored at –20°C.

SDS PAGE separating gel: 8-15% (w/v) 29:1 acrylamide:bis-acrylamide, 0.1% (w/v) SDS, 390mM Tris HCl pH8.8, 0.08% (v/v) TEMED, 0.1% (w/v) APS. Prepared immediately prior to use.

SDS PAGE stacking gel: 5% (w/v) 29:1 acrylamide:bis-acrylamide, 0.1% (w/v) SDS, 129mM Tris HCl pH6.8, 0.1% (v/v) TEMED, 0.1% (w/v) APS. Prepared immediately prior to use.

Tris-glycine electrophoresis buffer: 25mM Tris, 250mM glycine, 0.1% (w/v) SDS.

Bandshift binding buffer (5x): 30 mM Tris-HCl (pH8), 750 mM NaCl, 5 mM DTT, 30 mM MgCl2, 15% (v/v) Glycerol, 50 ng/µl BSA, and 0.05 µg/µl of poly(dAdT) (Amersham).

2.1.4 Bacterial Media

Ampicillin stock solution: 50mg/ml ampicillin in dH_2O. Filter sterilize (0.2µm filter) and stored at –20°C. Added to LB medium to a final concentration of 50µg/ml.

Kanamycin stock solution: 50mg/ml kanamycin in dH_2O. Filter sterilize (0.2µm filter) and stored at –20°C. Added to LB medium to a final concentration of 50µg/ml.

Choramphenicol stock solution: 34mg/ml chloramphenicol in 96% (v/v) ethanol. Stored at –20°C protected from light. Added to LB medium to a final concentration of 34µg/ml.

IPTG stock solution: 1M IPTG (Isopropyl-β-D-thiogalactoside) in dH_2O. Filter sterilize (0.2µm filter) and stored at –20°C.

X-gal stock solution: 40mg/ml X-gal in dimethylformamide (DMF). Protect from light and stored at –20°C.

Blue/white selection LB agar plates: LB / antibiotic plates were spread with 40µl 100mM IPTG, and 40µl 40mg/ml X-gal and dried at 37°C. Prepared on day of use.

LB medium: 10g/l Bacto tryptone (Difco), 5g/l Bacto yeast extract (Difco), 10g/l NaCl. pH adjusted to 7.0 with NaOH. 20g/l Bacto agar (Difco) added if making LB agar then autoclaved. LB agar plates (20ml volume) were stored inverted at 4°C and LB broth was stored at r/t.

NZY+ medium: 10g/l casein hydrolyaste, 5g/l yeast extract and 85mM NaCl. Adjust pH to 7.5 with NaOH and autoclaive. Prior to use add 12.5ml of 1M MgCl2, 12.5ml of 1M MgSO4 and 20ml 20% (w/v) glucose per 1l of autoclaived media.

Competent cell buffer A: 100mM RbCl, 50mM MnCl2, 30mM KOAc, 10mM CaCl2, 15%(v/v) Glycerol. Adjusted pH to 5.8 and filter sterilize (0.2µm filter).

Competent cell buffer B: 10mM MOPS, 10mM RbCl2, 75mM CaCl2, 15%(v/v) Glycerol. Adjusted pH to 8.8 and filter sterilize (0.2µm filter).

Miniprep solution 1: 25mM Tris HCl pH8.0, 10mM EDTA, 50mM glucose. Stored at 4°C.

Miniprep solution 2: 200mM NaOH, 1% (w/v) SDS. Prepared immediately prior to use.

2.1.5 Bacterial Strains

DH5α

Genotype: supE44 ΔlacU169 (ϕ80 lacZΔM15) hsdR17 recA1 endA1 gyrA96 thi-1 relA1. Laboratory stock.

XL1-Blue

Genotype: *recA*1 *endA*1 *gyrA*96 *thi*-1 *hsdR*17 *supE*44 *relA*1 *lac* [F′ *pro*AB *lac*I^qZΔM15 Tn*10* (Tet^r)]. Laboratory stock.

XL10-Gold

Genotype: Tetr Δ(mcrA)183 Δ(mcrCB-hsdSMR-mrr)173 endA1 supE44 thi-1 recA1 gyrA96 relA1 lac Hte [F′ proAB lacIqZΔM15 Tn10 (Tetr) Tn5 (Kanr) Amy]. Stratagene

TOP10F′

Genotype: F′{lacIq Tn10 (TetR) mcrA Δ(mrr-hsdRMS-mcrBC) Φ80lacZΔM15 ΔlacX74 recA1 araD139 Δ(ara-leu)7697 galU galK rpsL endA1 nupG. Invitrogen.

BL21-CodonPlus (DE3)-RIL

Genotype: E. coli B F– ompT hsdS(rB– mB–) dcm+ Tetr gal endA Hte [argU ileY leuWCamr]. Stratagene

BL21(DE3)pLysS

Genotype: E. coli B F– dcm ompT hsdS(rB– mB–) gal λ(DE3) [pLysS Camr]. Stratagene

StrataClone Solo Pack Competent Cells

Genotype: E. coli B F– lacZΔM15 [pSC-A-amp/kan] endA recA tonA. Stratagene

2.1.6 Microarray Reagents

MH buffer (Microarray hybridization): 2xSSC, 50% (v/v) deionised formamide, 10 mM Tris-HCl [pH7.5], 5% (w/v) dextran sulphate and 0.1% (v/v) Tween 20.

MHu buffer (microarray humidifier): 2xSSC, 50% (v/v) deionised formamide, 10 mM Tris-HCl [pH7.5] and 0.1% (v/v) Tween 20.

SSC (Saline Sodium Citrate - 20x): 3M Sodium Chloride (NaCl) and 0.3M Sodium Citrate. Adjusted to [pH7] using 1M HCl.

MWA buffer (Microarray Wash A): 1x PBS and 0.05% (v/v) Tween 20.

MWB buffer (Microarray Wash B): 1xSSC.

2.1.7 CXXC and MAP Affinity Purification reagents

BW1 buffer (Beads Wash 1): 20mM HEPES [pH7.9], 10% (v/v) Glycerol, 0.1% (v/v) TritonX-100, 100mM NaCl, 0.5mM PMSF and 10mM β-mercaptoethanol. β-mercaptoethanol and PMSF were added immediately prior to use. Filter sterilized (0.2μ), degassed and stored at 4°C.

BW2 buffer (Beads Wash 1): As for BW1 buffer + 10mM Imidazole.

CA buffer (Column buffer A): 20mM HEPES [pH7.9], 10% (v/v) Glycerol, 0.1% (v/v) TritonX-100, 0.5mM PMSF and 10mM β-mercaptoethanol. β-mercaptoethanol and PMSF were added immediately prior to use. Filter sterilized (0.2μ), degassed and stored at 4°C.

CB buffer (Column buffer B): As for CA buffer + 1M NaCl.

2.1.8 Oligonucleotides

All custom oligonucleotides were purchased from Sigma-Genosys. Oligonucleotides were resuspended in dH_2O to 100μM (except where stated). Working stocks were diluted in dH_2O to 10μM. Resuspended oligonucleotides were stored at –20°C.

Table 2.1-1. Genomic DNA PCR Primers

Seq ID	Primer Names	Sequence Forward	Sequence Reverse
NYESO	NYE' F2/R1	CCCAGCGTCTGGTAACCATC	CCACGGGACAGGTACCTC
p48	p48 F/R	CAGAAGGTCATCATCTGCCA	TGAGTTGTTTTTCATCAGTCCA
MAO_A	MAO F/R	CGGGTATCAGATTGAAACAT	CTCTAAGCATGGCTACACTACA
ATP SB	ATPSBS F/R	GAGGGTCTGGACGGCTGAGG	GGACTTCGGTGCTTACCTGG
cFOS	cFOS F/R	CTTCGGGAGGCAGGTTCGTTCT	GTTCCCGTTATCCCTTCAGCATC
XIST	XIST F2/R2	CACGTGACAAAAGCCATG	GGTTAGCATGGTGGTGGAC
Cathep'	Cath F/R	TGCCCCATAGACTCCAAGCCTCAG	TGTTCCCTCCCGCAAAGACTCA
hSex	H Sex-d F/R	CTGATGGTTGGCCTCAAGCCTGTG	TAAAGAGATTCATTAACTTGACTG
CXX1	104e15 F/R	CGGTGCGCATGCGCCAAG	CCGTCGGTTGCCGGATG
CXX2	23m08 F/R	GGCTGTTTTCGTTGGGAACG	GCGAATAAAAGCTGCTCGCG
CXX3	94k16 F/R	CATGTGATCTCAAAGAGGGC	CTCCGCTGTCCGGGAAG
CXX4	103c16 F/R	GGTTCTCTTGCCACTCAC	GAAGTGAAGCACCAGCG
CXX5	165h01 F/R	CAATGCAATGGTGAGGGTC	CCCTGCCCATTCGTTCAG
CXX6	65m08 F/R	GGTAGCCGTGACAGGTAC	GGGTGACCCAGGAGTTCG
CXX7	112 e08 F/R	CATTTAGCACTCCTGCCTCC	GGGAAAAAATACAAACCCTTC
CXX8	56a12 F/R	CCTTGATGGCCCCTTTTCAG	CATTGGAAAGGAACTGAGGC
CXX9	133l09 F/R	CCATGCTGGGAGATGACAC	CACTGTCTCCGATTCATC
CXX10	209f11F/R	CAACACTGATACTGAATAC	GTAATTCTTCCATGACTC
CXX11	49j09 F/R	GGCTCTTTCAGAGACTTC	GGCTCTCATTTCAAGGCC
CXX12	96d01 F/R	GCTCACACCCCTGCAAAC	GAGCTGGGTAGATACACTG
CXX13	267k21 F/R	CCCTTGTATTTTCCACAGC	GCAGAGTAAGCGGCTTTTG
CXX14	209n13 F/R	CAATATTCCGTAAGGAAAC	GCAGCAAAAAGTAAACTG
CXX15	92j24 F/R	GCTGGTAGTATTCATGTGTC	GATTTGGAAATGACTCCCAC
CXX16	275h22 F/R	CCTGAGGTACCCTAGAAGGC	TCCCTGGTCTGTCTGATGCC
CXX17	70p05 F/R	GGATCAGTCGCTATTGAGTG	TGGAGGAACATAACAGAATCG
CXX18	75n24 F/R	CCTCTGCTGTATGTCAGGTC	GTGCTCACTATGTGCCAGG
CXX19	213d20 F/R	GCATTATGCTAAATGCTGGG	GGAGTGGGGGAAGAAATGTG
CXX20	69a15 F/R	GCTAGGGGCTCTCAGATAAC	GGAACCAGAGTCTTGACTG
MAP 1	I2206 F/R	CGGTGAAAGAAGGAGGTGGG	CGTGACAAGAGAGTCCGCCT
MAP 3	I8192 F/R	CCTGGACCACACTCGTCCTA	GGACTTACTCCAGAAGGGCT
MAP 4	I8257 F/R	GCTCATTTCCACCGAGGTCAA	CGTAAGTGGGGACACTCATC
MAP 5	I10295 F/R	GGAACTTGGAGAGGAGGGCT	CCCAGCTCCTACTGTGGAAG
MAP 6	I10484 F/R	GGTTGTTTCAAGATGGCGGA	CCTCAACTTCGAGCTAGCTC
MAP 8	I16200 F/R	GCTAGAGATGGATGAGTCAC	GGGCAGTTCGGAATAAAAAC
MAP 9	I19422 F/R	GGGTTTCTTCTATTCCTCGG	GGCCTGTCACCCATTTGTTC
MAP 10	I122248F/R	CGTTCCTCTCTGCACTCAGG	GCATCCGTGAGGCAGATGTC
MAP 12	I6531 F/R	CGTCTTCGGCAGGTACTCAG	GCAGCCCTGATGCTCAACTG
MAP 13	I10840 F/R	GGGTTTATTTTGGTGGGACG	CCTCAGTTTCTCCAGTCAAG
MAP 14	I11855 F/R	GGCTAAAGAGCTGCTGCTGC	GGAGGCAGGTTCCTCTGTTG
MAP 16	I19049	CCATCCTCAGTTCCACCACC	GCAGGCCACGATAGATCAAC

Table 2.1-2. Bisulfite Primers

Seq ID	Primer	Sequence Forward	Sequence Reverse

	Names		
I3878	11b F3/R3	GATTGTAGTTAGTGAAATTGAAGTTAGA	AACCTAACCCCCAACCCTCCTAAAAAAC
I9112	17_2 F1/R1	GTAGGTATGATGTTAAAATTGAATTTGTG	ACCCCACCAAAAACAAAAAC
I2985A	I2985A F1R2	GGTATTTATTTATAAAAAAAGGG	TAAAAATCTCCCTATTCACTAC
I1878A	som12a F1/R1	ACTTATACTATCACATTTCCTTCCT	TTGTTTTTTAAGTGTTGTTAAGATTT
I1878B	som12b F2/R1	TATAGGAAGGAAATGTGATAGTATAAGT	ACAACTAAAACAAAAAACTCCTATC
I13406B	I13406B F2/R2	GTAGGTTAGTGATTAAGGTTTGTGT	CTCCTTCTAACCCCTAAACC
I11878	Som9 F1/R1	TTTTTAGATTTTGGGAATTTTAATA	CTCCAACAATAAAAACAAATCATC
I5134B	I5134B F2/R1	GAAGGTTGTGAGATTTTGGTTTAAT	CCTCCAAAACAAAACCTCTAAC
I12175	som15A F2/R1	GATGGTTAGTATAAAGTGTTAAT	TATAAACTTATAACTCTAACTAC
I3654A	I3654A F2/R1	GGTAGTTTTGTTATTTTGTGATAGTG	TTTCTAAAACAACCTTTCTTTTTCC
I3654B	I3654B F1/R1	GGAAAAAGAAAGGTTGTTTTAGAAA	AAACATCCAAATTAACACCATATATT
I3654C	I3654C F1/R2	AATATATGGTGTTAATTTGGATGTTT	CCAACACCTTATCCATCTATTTTA
I3360A	I3360A F2/R1	GTTTGTTTATTTGTTTGTTAATAAAGGG	CCCTACCCAACTTCCTCCAAAACTA

Table 2.1-3. Quantitative RT PCR primers

Seq ID	Primer Names	Sequence Forward	Sequence Reverse
G'PDH	GAPDH F1/R1 RT	TGGTATCGTGGAAGGACTCATGAC	ATGCCAGTGAGCTTCCCGTTCAGC
S'C31B	SEC31B FA/RA RT	CCACCTGAGAAGATGGAAAG	GCTTCTCATATAGATACTCCAG

Table 2.1-4. Sundry Oligonucleotides

Seq ID	Oligonucleotide Name	Sequence Forward	Sequence Reverse[a]
	Adaptor1 MseI	GGT CCA TCC AAC CGA TCT	1mM Stock
	Adaptor2 Mse1	(p)TAA GAT CGG TTG GAT GGA CC	1mM Stock
ImPCR	Universal Primer	GGT CCA TCC AAC CGA TCT TA	NA
MBD 77-167	MBD77-167 F/R	CGGTTCATAACCATATGGCTTCT GCCTCCCCCAAACAGCGG	CGGAAGTCAAAGAATTCTCATCAGTG GTGGTGGTGGTGGTGGTGCCGGGA

[a]Where applicable

2.2 Methods

All methods were carried out at R/T unless otherwise stated. Some standard techniques will not be outlined here.

2.2.1 DNA Manipulation

Human Samples and DNA extraction

Whole blood and sperm was collected from voluntary donors and anonymized, linked only with minimal sample information (Age, sex etc.). Donors were aware of, and consented to, its use for preparation of DNA.

Monocyte and granulocyte cells were prepared from whole human blood using Ficoll gradient centrifugation. Whole blood (3ml) was layered onto an equivalent volume of Histopaque-1077 Ficoll (Sigma-Aldrich) and sedimented at 400g for 30min according to manufacturers instructions. Granulocyte cell pellets were resuspended in 10mls of RCL buffer and centrifuged at 1200g for 10mins (repeated twice to remove red blood cells). Monocyte and granulocyte cells were rinsed and sedimented twice in PBS at 1200g. Cells were visualised using standard reverse field microscopy. Pelleted cells stored at -80°C until required. All centrifugation steps were carried out in a swing bucket Allegra X-22 Series Benchtop Centrifuge (Beckman Coulter).

Whole human blood, monocyte and granulocyte DNA was extracted using the Genomic-tip 500/G (Qiagen 10262) genomic DNA purification kit as described by the manufacturer.

Sperm DNA was prepared as described by Hossain and collegues (Hossain, et al., 1997). Briefly, 2ml total ejaculate was centrifuged at 5000g for 5min. Cell pellets were resuspended in 10ml PBS and sedimented as above with three repetitions to remove seminal material. Cell pellets were resuspended in SL buffer and incubated for 4hrs at 55°C. DNA was precipitated by addition of 20ml of Isopropanol (Propan-2-ol) and then spooled out into 70% ethanol. The DNA was air dried for ~5mins or until glassy in appearance and resuspended in 500µl of TE buffer.

Brain, Muslce and Spleen DNA were purchased (Ambion) and supplied at a concentration of 500µg/µl in TE.

Measurement of DNA concentration

DNA solutions were measured at OD_{260nm} and OD_{280nm} using a Nanodrop-1000 spectrophotometer. An automated reading of DNA concentration was calculated using Beer's law (Concentration_ng/ul = (Absorbance_OD_{260nm} x Extinction coefficient_dsDNA_50ng/µl/cm) / pathlength_cm). DNA purity was determined using OD_{260nm}:OD_{280nm} ratio, with ≥1.8 indicating the absence of residual protein or phenol contaminants from the sample.

Restriction digestion

DNA digest were carried out as per manufacturers instructions (NEB). Briefly, DNA was diluted in appropriate digestion buffer, supplemented with 100 µg/ml Bovine Serum Albumin where appropriate and digested with 6U of restriction endonuclease per µg of DNA. Reactions were typically incubated at 37°C for 1-2hours.

DNA Electrophoresis

DNA was resolved by agarose gel electrophoresis using the Sub-cell system (Bio-Rad). 0.8-2% (w/v) agarose gels were used depending on the DNA fragment(s) size to be resolved. Agarose gels were prepared with TAE (or TBE for fragments < 300bp) containing 0.5µg/ml ethidium bromide, a DNA intercalating agent that fluoresces in ultraviolet (UV) light. DNA samples and appropriate size ladder (100bp-12kb; Invitrogen) prepared in orange G loading buffer were loaded into the wells of the gel. Agarose gels were run at constant voltage (75-110V) in TAE (or TBE) electrophoresis buffer and visualised under UV light. Where electrophoresed DNA was to be further manipulated, UV exposure was kept to a minimum to prevent damage.

Gel Extraction

Gel extraction was used to purify a homogeneous population of DNA fragments for cloning or probe preparation. Fragments were resolved by agarose gel electrophoresis, cut out and extracted using the Perfectprep Gel Cleanup Kit (Eppendorf) according to manufacturers' instructions.

In vitro methylation

DNA was methylated using the cytosine methyltransferase M.SssI (NEB). Methylation was carried out in a volume of 200µl containing 0.5-2µg, DNA, 2µl M.SssI (8U), reaction buffer 2 (NEB), and supplemented with 160µM SAM. After an ON incubation at 37°C the reaction was supplemented with a further 2µl M.SssI and 1 µl SAM (32mM) and incubated for 2hrs prior to reaction clean up.

End labeling

DNA was radiolabelled by klenow incorporation of α^{32}P-dCTP. Labeling was carried out in a total volume of 50 µl containing 200-800ng DNA, 5µl 10mM dNTPS (-dCTP), 2 µl α^{32}P-dCTP (GE Healthcare), 2ul klenow large fragment (NEB) and reaction buffer 2 (NEB). The labeling reaction was incubated at 37°C for 2 hours and then DNA was extracted using a nucleotide clean up kit (Qiagen).

Oligo Annealing

Oligos A1 and A2 at a concentration of 1mM were annealed by incubation at 100°C for 10mins and then cooling for 5 hours to R/T. Final adaptor concentration of 500µM.

DNA ligation

Standard DNA ligation reactions were performed in a total of 20µls containing 100ng linearised vector, insert DNA (3x molar excess), T4 DNA ligase buffer and 1ul T4 DNA ligase (NEB; equivalent to 6 Weiss units) and incubated ON at 16°C.

CGI library cloning was performed in a total volume of 1.2ml containing 1/3rd of the CXXC purified CGI fraction, 600ng pGEM5zf-(gel extracted), 40µl of T4 DNA ligase (240 Weiss units) and T4 DNA ligase buffer. The ligation reaction was incubated at 16°C for 24hours.

MAP array linker ligation for lmPCR. 10µl universal adaptor (500µM), 8µl T4 DNA ligase (48 Weiss Units) and T4 DNA ligase buffer was added to 50µg of MseI digested genomic DNA to a total volume of 500µl and incubated at 16°C ON. A further 4 hour incubation was carried out after the addition of 4ul of T4 DNA ligase (24 Weiss units). Ligation reactions were cleaned up using a PCR clean up kit (Qiagen) according to manufacturer's instructions.

DNA sequencing

DNA sequencing was performed using the BigDye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems). Sequencing reactions contained 2µl BigDye Terminator v3.1, 5pmol primer, 4µl DNA sequencing buffer and 10-150ng DNA template in a 10µl volume. The reactions were incubated under the following conditions: initial DNA denaturation at 96°C for 10secs, then 24 cycles of DNA denaturation at 96°C for 30s, primer annealing at 50°C for 20s, and extension at 60°C for 4min. Sequencing reactions were cleaned up and were run on an ABI 3730 capillary sequencer by the School of Biological Sciences Sequencing Service. Sequence data was quality controlled and analysed using Chromas Lite and Lasergene software respectively.

The CGI library was sequenced at the Wellcome Trust Sanger Institute in Hinxton, Cambridge. The clone set was arrayed into 384 well plate format and sequenced using a modified alkaline lysis method. Cells were lysed in glucose, Tris, EDTA (pH 8) buffer plus NaOH and SDS and spun through Millipore Montage filter plates followed by DNA precipitation and resuspension in dH_2O. In all 172,800 clones were sequenced forward and reverse using T7 and SP6 primers and BigDye V3.1 chemistry, under the following conditions: 30sec at 96°C, followed by 44 cycles of 92°C 8sec, 55°C 8sec, 60°C 2min. Samples were resolved using a 3730 XL sequencers (Applied Biosystems). Extraction was performed using sequence analysis v3.1, and base-called using Phred. DNA sequences were identified and mapped to the human NCBI build 36 and annotated on the ENSMBL Genome Browser (http://www.ensembl.org/index.html).

Polymerase Chain Reaction (PCR)

PCR was employed to amplify specific target DNA molecules from a DNA template. Typical reactions were 25µl, containing DNA template (0.1-100ng), 400nM forward and reverse primers, 400µM dNTPS, Red Hot PCR buffer (ABgene), 2.5mM $MgCl_2$, 1.5U Red Hot DNA polymerase (ABgene). A negative control reaction with dH_2O and no template and a positive control with template and conditions previously shown to yield a product were included. PCR cycling conditions were as follows: an initial denaturation at 94°C for 2min followed by 30 cycles of denaturation at 94°C for 40sec, primer annealing at T^{ann} for 50sec, primer extension at 72°C for 50sec and an additional 72°C primer extension phase for 7min to amplify any incomplete DNA molecules. Specific primer annealing temperatures (T^{ann}) were generally calculated to be the lowest primer melting temperature minus 2°C.

Depending on the purpose of the PCR some or all of the product was resolved by agarose gel electrophoresis. Supplementing the PCR reactions with 3% DMSO or Betaine or increasing the $MgCl_2$ concentration was routinely employed to amplify very GC rich target sequences.

Whole genome amplification for microarray labelling was achieved using linker mediated (lm)PCR. Reactions were carried out in a volume of 50µl containing adaptor ligated DNA template (10-100ng), 400uM dNTPS, 500µM $MgCl_2$ supplement (GC Rich PCR system; Roche), 400nM lmPrimer, 5 µl GC Rich supplement (GC Rich PCR system - Roche), reaction buffer (GC Rich PCR system - Roche) and 0.5 µl polymerase (2.5U; GC Rich PCR system - Roche). PCR cycling conditions were as follows: an initial denaturation at 95°C for 3min followed by 20 cycles of denaturation at 95°C for 50sec, primer annealing at 58°C for 50sec, primer extension at 72°C for 4min and an additional 72°C primer extension phase for 7min. Amplified DNA was purified using a PCR clean up kit (Qiagen).

Quantitative PCR

Quantitative PCR was used to accurately quantify a target region in a DNA sample. SYBR Green I is a cyanine dye that absorbs light at a λ_{max} of 488 nm and emits light at a λ_{max} of 522 nm when complexed with double stranded (ds) DNA. Measurement of emitted light (522nm) directly allows the real-time quantification of duplexed DNA as it is synthesised. Each PCR reaction was prepared in a total volume of 20 µl, including DNA template (1-50ng), SYBR Green supermix (hot start i*Taq*, sybr green dye and dNTPS; Biorad) and 300nM forward and reverse primers. Reactions were cycled on an iCycler iQ (Bio-Rad) according to manufacturer's instructions. Cycling conditions were as follows: an initial denaturation step of 95°C for 2.5mins then 45 cycles of denaturation at 95°C for 30sec, primer annealing at T^{ann} for 30sec, primer extension at 72°C for 30sec and data collection. After cycling, a single denaturation step at 95°C for 1min is followed by a 35°C to 94.5°C ramp spanning 20min in 119 in 0.5°C incremental steps. The final temperature range generates the product melting curves.

iCycler iQ software assigns a baseline fluorescence measurement based on the standard deviations calculated for cycles 2-10. This baseline bisects the fluorescence curves from each PCR in the linear amplification phase and indicates the cycle threshold value for that reaction (C_T). An arbitrary measure of DNA quantity (Q) can be calculated using:

$Q = 2^{-Ct}$

Target sequences were quantified within and between DNA samples when calculated as a ratio to a reference sequence or sample. In order to ensure that quantifications were significant each PCR was carried out in triplicate and standard deviations were calculated for each quantification.

Cloning

DNA sequence corresponding to the CXXC and MBD domains were introduced into the pet30b His-tagging vector (Novagen). Briefly, DNA sequence corresponding to the hMBD and mCXXC domains were PCR amplified from cDNA. Forward and reverse primers complementary to the coding sequence, included an NdeI and EcoRI site respectively. NdeI introduces an ATG translation start site in frame with the cloned coding sequence. To incorporate a Histidine tag, the forward (N terminal) or reverse (C terminal) primer included an additional 18 nucleotides coding the 6xHis. Gel extracted PCR products were digested with EcoRI(NEB) and NdeI(NEB) according to the manufacturers instructions and ligated into the corresponding sites in the MCS of pet30b.

PCR fragments generated by Taq DNA polymerase were gel extracted and cloned using the StrataClone PCR Cloning Kit (Stratagene) according to manufacturers' instructions. Abundant PCR products were cloned using half reactions including only 50% of the template, vector mix, buffer, competent cells and bacterial growth media. All incubation steps were carried out according to manufacturers' instructions.

Exonuclease I/Phosphatase treatment

To prepare PCRs for sequencing 0.25µl (5U) exonuclease I (NEB), 1µl (5U) Antarctic phosphatase (NEB) and 11.75µl of dH_2O was added to 15µl of PCR product. Exonuclease degrades ssDNA (single stranded DNA) primers and phosphatase removes the phosphate groups from the dNTPS preventing these residual PCR components from interfering with the subsequent sequencing reaction.

Bisulfite Genomic Sequencing

Genomic DNA (2-5µg) was digested with a restriction endonuclease (cleaved outside the region of interest) and precipitated. DNA was resuspended in 25µl TE buffer, transferred to a siliconised tube and denatured at 100°C for 5min before returning to ice. To aid denaturation, 2.5µl of freshly prepared 3M NaOH was added to the DNA and incubated at

37°C for 20mins. 270µl of bisulfite modification solution was added to the DNA then overlayed with 200µl of mineral oil. The bisulfite reaction was incubated at 55°C for 5hours protected from light prior to isopropanol precipitation (including 50µg glycogen carrier; Roche). DNA was resuspended in 25µl and desulfonated by the addition of 2.5µl of freshly prepared 3M NaOH and incubation at 37°C for 15min. The DNA was EtOH / NH_4OAC precipitated and resuspended in 40µl of TE.

2µl bisulfite-treated DNA was used in a standard Red Hot PCR reaction supplemented with 3% DMSO and containing primers specific for bisulfite-converted DNA. PCRs were resolved by agarose gel electrophoresis (TAE), gel extracted and PCR cloned. The cloning reaction was transformed into StrataClone Solo Pack Competent Cells (Stratagene) according to the manufacturer's instructions. Transformed cells were spread onto blue/white selection ampicillin lb agar plates. Colonies harboring plasmids containing inserts were identified by blue/white screening (insertion into the cloning site disrupts the α-fragment of β-galactosidase). Inserts were PCR amplified from single colonies by transferring some cells with a pipette tip into a 20µl Red Hot PCR reaction containing primers flanking the vector insert site. 5µl of each PCR was resolved by agarose gel electrophoresis to confirm that cloned fragments were of the correct size. PCRs corresponding to validated insert were Exonuclease I/Phosphatase treated and then sequenced using one of the flanking primers from the colony PCR.

Bisulfite genomic sequencing data was analysed using the BiQ_Analyzer software package (Bock et al., 2005). Much of the cloning, colony selection and DNA sequencing was carried out by Dina DeSousa and Elaine Evans.

Cyanine Labeling

Florescent labeling of DNA for microarray hybridization. Bioprime random primer mix (Invitrogen) was added to DNA samples (100-300ng) to a final volume of 130.5µl, denatured at 100°C for 10min and then immediately chilled on ice. On ice, 15µl of dNTP mix (1 mM dCTP, 2 mM dATP, dTTP and dGTP), 1.5µl Cyanine 3 or 5-dCTP (depending on dye swap orientation; GE Healthcare) and 3µl of klenow (120U; Invitrogen) was added to each reaction. These were incubated at 37°C for 4hours (protected from light). Labelled DNA was purified using the Purelink PCR cleanup kit (Invitrogen) according to manufacturers instructions and elute in 100ul elution buffer. Cyanine incorporation was

calculated using a Nanodrop spectrophotometer and labeled samples were stored short term at -20°C.

2.2.2 RNA Manipulation

RNA extraction

Monocyte and Granulocyte RNA was extracted using Tri reagent (Sigma-aldrich). Tri reagent contains phenol and guanidine isothiocyanate and serves to maintain RNA integrity during cell disruption and dissolution of cell components. Addition of chloroform and centrifugation allows biphasic separation such that RNA may be retrieved from the aqueous phase whilst DNA and protein remains in the organic phase. Monocyte and Granulocyte cell preparations from 3mls of whole blood (prepared as outlined in Human samples and DNA extraction) were resuspended in 1ml of Tri reagent (Sigma-aldrich) and incubated at r/t for 5min to disrupt nucleoprotein complexes. Following the removal of insoluble material by centrifugation (10min at 12,000g at 4°C) 200µl chloroform was added, mixed vigorously and incubated for 10min at R/T. The aqueous phase (RNA) was resolved from the organic phase (DNA and proteins) by centrifugation for 15min at 12,000g at 4°C. To precipitate the RNA the aqueous phase was transferred to a tube containing 500µl of isopropanol, mixed and allowed to stand for 10min at r/t. RNA was pelleted by centrifugation (10min at 12000g at 4°C), washed with 70% EtOH, air dried and resuspended in 50µl of nuclease free water (Ambion). RNA was resolved by RNA gel electrophoresis (see RNA electrophoresis) to address the quality of the preparation as indicated by sharp bands representing the 28S and 18S ribosomal RNA. RNA was stored until required at -20°C.

cDNA synthesis

Prior to cDNA synthesis, contaminating DNA must be removed from the RNA preparation to avoid artefactual results. RNA (~500ng) was treated in a total volume of 40µl containing 3ul (6U) DNAse I (DNA-free kit; Ambion) in DNAse I reaction buffer (DNA-free kit; Ambion) and incubated for 30min at 37°C. DNase I was inactivated according to the manufacturers instructions.

Prior to reverse transcription the RNA was denatured at 65°C for 5min and snap chilled on ice. 250ng of the RNA was prepared in a total volume of 25µl containing 5µM random hexamers (Roche), 1mM dNTPs (Abgene), 40U RNasin (RNase inhibitor; Promega), 200U

of M-MLV reverse transcriptase (RNase H minus; Promega) and MLV reverse transcriptase reaction buffer (Promega). In addition a minus (-)RT reaction was set up in parallel containing all the components except for the reverse transcriptase to control for DNA contamination by PCR. The reactions were cycled 4x each each for 8min at 20°C, 8min at 25°C and 30min at 37°C followed by final heat inactivation at 70°C for 15min. The 25µl RT reactions were stored at –20°C. 1µl of cDNA was used in each PCR in parallel with the respective (-)RT control.

RNA Electrophoresis

RNA was resolved using the Sub-cell system (Bio-Rad). To remove ribonucleases, gel apparatus was washed with 0.4M NaOH and thoroughly rinsed with filtered dH_2O prior to electrophoresis. RNA samples and size marker (281bp-6.58kb; Promega) were denatured in RNA loading buffer for 10min at 70°C and chilled on ice. These were then loaded onto a 1.5% (w/v) non denaturing agarose gel containing 0.5µg/ml ethidium bromide. Agarose gels were run at constant voltage (80V) in TAE buffer and visualised under UV light.

Quantitative PCR

Relative RNA transcript levels were determined using quantitative PCR essentially as described for DNA. Each PCR contained 1µl of cDNA or (-)RT except when quantifying an abundant transcript such as hGAPDH when the starting material was diluted by 1:10.

2.2.3 Protein Manipulation

Preparation of bacterial whole cell lysate

Bacterial cells were pelleted at 4200rpm for 30min at 4°C and washed with cold PBS. The cell/PBS suspension was then pelleted at 1l starting culture per tube and stored at -80°C. The bacterial pellets were resuspended in cold NAL buffer containing 1mg/ml lysozyme (disrupts the bacterial cell wall) and incubated on ice for 30min. The lysates were sonicated (Branson sonifier 250) at a duty cycle of 30% and an output setting of 5 for 5min on ice to sheer the bacterial DNA and aid downstream purification. The lysate was then centrifuged at 17,000g for 40min at 4°C and the clarified soluble material transferred to a clean tube and stored at -80°C.

Nickel affinity chromatography

All stages of this protocol were carried out at 4°C unless otherwise stated. Nickel beads were prepared by incubating chelating sepharose FF(GE Healthcare) in 0.2M $NiSO_4$ for 10min. The charged beads were washed with 5 volumes of dH_2O, 5 volumes of chelating sepharose wash buffer, rinsed with dH_2O and stored until required in 20%(v/v) EtOH. Nickel beads were equilibrated with 2 volumes of cold N10 buffer and added to the clarified bacterial lysate at 0.5ml of beads per litre of initial bacterial culture. The beads were agitated for 2hours to ensure maximal protein adsorption and then pelleted at 500g for 5min after which the unbound supernatant was removed. The beads were washed twice with 10 bed volumes of cold N20 buffer. A third wash was transferred to a disposable chromatography column (GE Healthcare) and allowed to flow through the column. Tagged protein was eluted from the column by competition with the histidine analogue Imidazole. One bed volume of N250 buffer was added to the beads and equilibrated for 10min and then allowed to flow from the column. This process was repeated 7 times. Samples of the soluble lysate, unbound material, wash and eluted fractions were resolved by SDS PAGE to assess the efficiency of the purification procedure. Fractions containing appreciable quantities of purified proteins were pooled and dialyzed in a slidealyzer cassette (7kDa cut off; Pierce) overnight in 2x2litres cold NAL buffer to remove imidazole.

Cation Exchange Chromatography

Nickel affinity purified protein was dialyzed in a slidealyzer cassette (7kDa cut off; Pierce) overnight in 2x2litres of 100mM NaCl cation exchange buffer (90%CEA and 10%CEB). The protein was bound to a Mono-S strong cation exchange column (Methyl sulfonate active groups; GE Healthcare) and eluted into 1ml fractions spanning a 0.1-1M NaCl gradient using an AKTA purifier liquid chromatography system (GE Healthcare). Samples of the input, unbound material, wash and eluted fractions were resolved by SDS PAGE to assess the efficiency of the purification procedure. Fractions containing appreciable quantities of purified proteins were pooled, dialyzed into cold NAL buffer and stored at -80°C.

Coupling Protein to Cyanogen Bromide Sepharose beads (CNBr beads)

All stages of this protocol were carried out at 4°C unless otherwise stated. 500mg of lyophilised CNBR-activated sepharose (GE Healthcare) was resuspended in 5ml of 1mM HCl (pH3; 500mg equivalent to 1.75ml of swollen media). The beads were transferred to a disposable chromatography column (Econo-Pac; Biorad) and washed with 100 bed volumes of 1mM HCl (pH3). CNBr beads were then equilibrated with 10 bed volumes of cold CNBr coupling buffer. 40mg of purified recombinant protein was dialyzed in a slidealyzer cassette

(7kDa cut off; Pierce) overnight in 2x2litres cold CNBr coupling buffer. Protein was added to 1ml of pre-equilibrated media and coupled for 2hours with agitation. Samples of input and unbound protein were measured by Bradford assay to assess the coupling efficiency. Coupled beads were then rinsed in 90%CA / 10% CB buffer prior to column preparation.

Protein Concentration measurement

The Bio-Rad Protein Assay was used to determine the concentration of protein preparations. The assay works on the principle that the λ_{max} of Brilliant Blue G-250 dye shifts from 465 nm to 595 nm when complexed with protein. A range of protein standards (lysozyme or BSA) were incubated with a 1:5 dilution of the dye reagent (Biorad) for 5mins and the OD_{595} measurements used to plot a standard curve. Unknown protein concentrations were determined using OD_{595} measurements substituted into the equation describing the standard curve.

Bandshift Assay

A Bandshift assay (also gel shift of EMSA) determines the specificity of protein DNA binding. Radiolabeled DNA probes are incubated in the presence of protein and then resolved by gel electrophoresis. Protein:DNA binding is indicated by reduced mobility through the gel matrix. Recombinant MBD and CXXC (0-1000ng) was incubated with 0-2µg poly(dA.dT) (non-specific competitor DNA; Sigma) and binding buffer at r/t for 10min. 15pmol of radiolabeled methylated or nonmethylated CG11 probe and 0.1%(w/v) bromophenol blue was added to the reactions and incubated for a further 25min at r/t. The bandshift reactions were loaded onto a pre-cooled 1.3%(w/v) 0.5x agarose TBE gel and run at 120V for 3hours. Bandshift gels were placed onto 2 layers of DE81(Whatman) and 2 layers of 3M Whatman (VWR) paper and dried at 80°C for 2hours using a speed vac. Dried gels were incubated in a Phosphor screen for 2-5hours and imaged using a Storm 860 scanner (GE Healthcare) and analysed using ImageQuant software (GE Healthcare).

Protein Electrophoresis

SDS polyacrylamide gel electrophoresis (SDS PAGE) was used to resolve proteins based on molecular weight. Both the Mini-PROTEAN 3 system (Bio-Rad) and Sci-Plas system (Sci-Plas) were utilised depending on the application. Stacking gels contained 5% (w/v) acrylamide and separating gels between 8 to 15% (w/v) acrylamide depending on the molecular weight of the proteins of interest. Protein samples, prepared in protein loading buffer, and Pre-stained broad range protein markers (6-175kD; NEB) were incubated at

100°C for 5min to aid SDS coating and to disrupt quaternary structures. Samples were then chilled on ice and loaded into the wells of the stacking gel. Electrophoresis was carried out at 25mA (Mini-PROTEAN 3 system) or 65mA (Sci-Plas system) in Tris-glycine electrophoresis buffer.

SDS PAGE gels were Coomassie stained by incubation in Coomassie Brilliant Blue R-250 staining solution with agitation for 20min. Gels were de-stained by immersion in Coomassie destain solution until background was reduced sufficiently. A final rinse in dH_2O at 100°C was included to remove the last traces of non-specific staining. Gels were scanned and then dried between two sheets of clear cellulose film (Biorad).

2.2.4 Bacterial Preparation

Bacterial Conditions

Bacterial transformations were streaked onto LB agar plates containing the appropriate antibiotics, inverted and incubated at 37 °C overnight. Single colonies ere picked with a sterile loop and transferred to LB broth with appropriate antibiotics and incubated at 37 °C overnight.

Competent Cells

5ml of LB culture was grown overnight at 37°C and then diluted 1:100 into a final volume of 500ml. The culture was incubated at 37°C until reaching an OD_{600} of 0.5 and transferred to 2x sterile prechilled centrifugation bottles and incubated on ice for 10min. Cells were pelleted at 2400rpm for 15 at 4°C and then resuspended in 165ml of cold competent cell buffer A. After incubation on ice for 45min the cells were pelleted as before and resuspended in 40ml of cold competent cell buffer B. Cells were incubated on ice for 15min and then snap frozen in liquid N_2 in 200µl aliquots. Competency was calculated through a transformation titration of 0,1,10 and100ng of plasmid DNA. Efficient competent cells should yield $>10^6$ colonies per µg of plasmid DNA.

Transformation

Calcium chloride competent cells can be heat shocked to induce uptake of naked DNA molecules. 10-100ng of plasmid DNA or 2µl of a ligation reaction was incubated with 100µl of competent cells for 25min on ice. The cells were incubated at 42°C for 45sec and returned

to ice for 2min. 900µl of 37°C NZY+ media was added to the transformation and were then incubated with agitation for 1hour at 37°C. Depending on the starting quantity of plasmid DNA and the expected efficiency of transformation, 25-500µl of the cells were plated on an appropriate lb/antibiotic plate, inverted and incubated at 37°C overnight.

Bacterial Expression of recombinant proteins

The Bl21 E. coli strain and its derivatives were used for over-expression of target genes under the control of a T7 promoter. Bl21 cells and the pet30b cloning vector (Novagen) both encode the lacI repressor protein and Bl21 also encodes bacteriophage T7 RNA polymerase under the control of the lac promoter/lacO system. On addition of IPTG to the culture media, the lacI repressor protein complexes with the IPTG and prevents binding to the lac operator. This allows expression of T7 RNA polymerase and in turn induces the expression of the recombinant protein from the T7 promoter. The CXXC construct was expressed in Bl21 pLysS which expresses nominal levels of T7 lysozyme. This binds to and inhibits T7 RNA polymerase preventing leaky expression of cytotoxic proteins.

An overnight culture of Bl21 transformed with the desired construct was diluted 1:100 into lb media and the appropriate antibiotics. The culture was then incubated at 37°C until reaching an OD_{600} of 0.45-0.6. 1mM IPTG was added to the media to induce recombinant protein expression and was incubated for a further 2-3hours. When expression of a protein prone to degredation, induction was carried out at 30°C for 5hrs.

Isolation of plasmid DNA

Plasmid DNA was prepared from Xl10 Gold, DH5α and Xl1 Blue E.coli strains. Each of these strains is endonuclease deficient (*endA*) and recombination deficient (*recA*) which helps stable maintenance of the insert sequence.

For general plasmid preparations for applications such as probe production or restriction digestion the following protocol was used. 2ml o/n bacterial culture pelleted at 12,000g for 1min, and the cell pellet was resuspended in 100µl Miniprep solution 1 with 100µg/ml RNase A. The cells were then lysed by the addition of 200µl Miniprep solution 2, and then neutralised with the addition of 150µl 5M KOAc pH4.8. Insoluble cell debris were removed by centrifugation at 14,000rpm for 5min. Plasmid DNA was then recovered from the

supernatant by ethanol precipitation, and was resuspended in 50µl TE buffer. Plasmid DNA was stored at –20°C.

Where plasmid DNA was to be used for sequencing, cloning or if larger quantities were required, DNA was prepared using plasmid isolation kits (Qiagen) according to manufacturers instructions.

2.2.5 CAP and MAP Purification

Preparation of MBD and CXXC Chromatography columns

50-60mg of recombinant CXXC or MBD was dialyzed into NAL buffer and bound to 1ml of nickel charged sepharose at 4°C for 2 hours. The beads were washed with 10 column volumes (cvs) of BW1 buffer, 10 cvs of BW2 buffer and 10CVs of BW1 buffer. The beads were packed onto a 1ml chromatography column (Tricorn 5/50; GE Healthcare) at a constant 1ml/min linear flow rate. Once assembled, the chromatography column was equilibrated with a 200ml linear NaCl gradient (0.1-1M; gradient formed between CA and CB buffers) at a 1ml/min flow rate on an AKTA purifier (GE Healthcare). This equilibration serves to pack the beads and remove any bacterial DNA from the matrix.

CAP and MAP Chromatography

All chromatography steps were carried out on either an FPLC or an AKTApurifier (GE Healthcare) at a linear flow rate of 1ml/min and a maximum backpressure of 0.3MPa. Fragmented DNA (25-100µg) was EtOH precipitated and resuspended in 500µl of 100mM NaCl column buffer (90% CA and 10% CB). DNA was bound to the chromatography column by injection. Bound DNA was then eluted over an increasing NaCl gradient of 0.1-1M by mixing CA and CB buffers. 3ml fractions were collected across the gradient. 250 µl of each fraction was precipitated and resuspended in 40µl TE buffer for fractionation analysis.

2.2.6 Microarray Procedures:

Microarray experiments were carried out using custom designed arrays generated by Cordellia Langford and colleagues at the Wellcome Trust Sanger Institute. CGI amplicons were amplified from the CXXC library to include 5' primary amine groups. These were coupled to amine binding slides and denatured to remove non-coupled DNA molecules. The resulting array consists of 28.8k features representing of 17,387 unique single stranded CGIs.

Hybridization and washing

Cyanine labeled MAP and Input DNA and 100µg of human Cot-1 DNA (Invitrogen) were EtOH precipitated (protected from light). In parallel a prehybridisation mix containing 100µg of human Cot-1 DNA (Invitrogen) and 40µg of Sonicated herring sperm DNA (Sigma-Aldrich) was prepared and precipitated. Both the hybridization and prehybridisation mixtures were resuspended in 60µl of MH buffer, denatured at 100°C for 10mins and snap chilled on ice.

The prehybridization was then incubated at 70°C for 10mins and applied the surface of a pre-chilled coverslip. A prechilled microarray was lowered onto the coverslip until adhesion with the hybridisation mix lifted it onto the array surface. Once the coverslip was orientated across the full arrayed area and any air bubbles had been expelled, the array was transferred to a humidified chamber containing 1ml MHu buffer and incubated for 1 hour at 37°C. The coverslip was then carefully removed by submerging the array in MWA buffer. The array was washed twice in MWA buffer for 10min, one in MWB buffer and rinsed twice in dH2O. Microarrays were centrifuge dried at 500g for 1min in 50ml disposable tubes such that the ID label was at the bottom of the tube.

The hybridization mix was incubated for 1 hour at 37°C and then applied to the microarray as for the prehybridisation. Hybridisation was carried out for 48hrs at 37°C. Once the coverslip was removed the microarrays were washed four times for 10min in 37°C MWA buffer, three times for 10min in 52°C MWB buffer, twice for 10min in r/t MWA and finally rinsed twice in dH2O, before being dried by centrifugation (500g). Hybridised microarrays were protected from light and moisture and scanned within 24 hours.

Scanning and Data acquisition

Arrays were scanned with a 2 laser GenePix Autoloader 4200AL (Axon). Arrays were scanned such that approximately ~5% of features were saturated. Microarray scans representing Cyanine 3 and 5 signals were processed using the GenePix Pro 6.0 (Axon) software package. Prior to calculating the signal intensity values, all spots were manually checked and flagged for later analysis.

2.2.7 Bioinformatic Analysis

Microarray analysis, figure generation, and Gene Ontology (GO) analysis weas carried out by R Illingworth. The majority of gene mapping and genome wide database searching was carried out by the WTCCB Bioinformatician, Dr Alastair Kerr.

Microarray analysis

Analysis was carried out with the LIMMA package in the R statistical environment. Flagged CGIs were removed from the analysis and features with poor signal-to-noise ratios were stabilised using a minimum capped value of 1000 for background subtracted intensities. Cy3 and Cy5 signals were transformed into M values (log2 [red/green]) and normalised by print-tip loess. Each biological variable analysis was represented by four microarrays comprising two independent replicates with two respective dye swap hybridisations. Processed values were averaged through linear modeling and used to determine the relative enrichment of MAP DNA relative to Input. An M value of ≥ 1.5 was designated as the threshold for hypermethylation as determined by Quantitative PCR and bisulfite genomic sequencing. This threshold was confirmed as significant by calculation of a t-statistic by Bayesian modeling and BH multiple testing correction. Differential methylation was deduced when features displayed an M value ≥ 1.5 in one or more tissues and a differential of 0.75 between tissues (upper boundary capped at M = 2.5). CGIs which failed to satisfy these criteria but which had an M value ≥ 1.5 in one or more tissues were classified as 'uncharacterised' as it was not possible to conclude the methylation status in the remaining tissues. To avoid complications due to X chromosome inactivation, CGIs on sex chromosomes were excluded from the analysis of tissue specific methylation. In addition, spots that gave no signal on the microarray (NA values) and spots containing DNA in which CpG[o/e] values were <0.5 were excluded. All microarray data is available in MIAME compliant format in Array Express (http://www.ebi.ac.uk/microarray-as/ae/) accession number: E-MEXP-1391 (01/07/2008).

Significance testing

Data distribution was determined using a Shapiro-Wilk test of normality. Data conforming to a parametric distribution were analysed using a Welch Two Sample t-test. For non parametric data sets, significance was determined using either a two sample Kolmogorov-Smirnov tests or a Wilcoxon rank sum test as indicated.

Genome mapping and data mining

Physical mapping of sequences to the human genome was carried out by Dr Alastair Kerr. CGIs that mapped within 1.5kb of annotated genes were considered to be gene-associated. This window takes into account mis-annotation of transcription start sites within poorly defined 5'UTRs. Genic association was broken down into 4 categories. Promoter associated was designated as +/- 1.5kb from the most 5' end of an annotated gene. 3' associated was designated as +/- 1.5kb from the most 3' end of an annotated gene. Intragenic was designated as overlapping an annotated gene internal to the most 5 and 3 extremities whilst not being within 1.5kb of them. Intergenic was designated as not within 1.5kb of an annotated gene. Where multiple alternative transcripts were reported for a given gene the largest was used for analysis.

Chapter 3: CXXC Affinity Purification

3.1 Introduction

3.1.1 CpG Islands

The methylated, CpG deficient, mammalian genome is punctuated by discrete nonmethylated DNA sequences called CGIs. These G+C rich islands are approximately 1 kilobase in length and show little CpG suppression relative to bulk genomic DNA. CGIs co-localize with approximately 56% of human gene promoters and as such are overrepresented in R bands (Antequera and Bird, 1993; Craig and Bickmore, 1994; Larsen et al., 1992).

Despite 25 years of investigation the exact functional role of CGIs remains unclear. Transfection studies have shown that transcription from a CpG island promoter can be extinguished with the introduction of DNA methylation (Stein et al., 1982). Conversely, genome wide demethylation induced by the drug 5AzaC has been shown to activate CGI genes that were previously methylated (Hansen and Gartler, 1990). Aberrant CGI methylation associated with neoplastic cells is also known to correlate with gene silencing, although it is unclear whether this is causal. Nonmethylated CGIs are generally transcriptionally permissive and associate with all constitutively expressed 'housekeeping genes' (Larsen et al., 1992). Accordingly, mutagenesis of Sp1 binding sites from the mouse *APRT* promoter CGI results in *de novo* methylation and transcriptional repression (Brandeis et al., 1994; Macleod et al., 1994). In addition to gene promoter association, several mammalian CGIs have been identified as initiation sites of DNA replication (Delgado et al., 1998; Phi-van and Stratling, 1999). This is consistent with elevated G+C levels which is not solely accounted for by lack of CpG suppression within these regions (Antequera, 2003; Antequera and Bird, 1999). As such, the distinct sequence characteristics of CGIs represent a unique genomic landmark, which are of significant value for gene mapping and the elucidation of genome function.

3.1.2 The number of CGIs in the human genome

CGIs were initially discovered by restriction digestion of vertebrate genomic DNA with the methyl-sensitive restriction endonucleases Hpa11 (CCGG). Cleavage patterns indicated two

distinct genomic fractions based on CpG density and DNA methylation status. High molecular weight DNA, refractive to digestion, was methylated and CpG deficient. The second fraction, composed of very highly fragmented DNA, represented a small proportion of the genome which was nonmethylated and CpG rich (HTFs; (Bird et al., 1985; Cooper et al., 1983)). HTFs were found to originate from 1kb regions enriched at gene promoters and were subsequently renamed CpG islands (CGIs) to better suit their sequence composition (Bird, 1987; Gardiner-Garden and Frommer, 1987). This procedure was quantified to allow the prediction of the total number of CGIs, estimating haploid complements of 44,000 and 37,000 in human and mouse respectively (Antequera and Bird, 1993; Bird et al., 1985). The human figure was later corrected for bulk genomic DNA contamination to suggest a more modest 26,300 CGIs (Ewing and Green, 2000).

As genome sequence has become available, CGI identification has moved increasingly toward *in silico* prediction methods. The first sequence prediction effort utilized CpG island sequences associated with promoters annotated in GenBank. Characterization of these sequences led to the first purely sequence based description of CGIs, stating that they must be at least 200bp in length, 50% G+C and have a minimum CpG[o/e] density of 0.6 (Gardiner-Garden and Frommer, 1987). This study predates the completion of the human genome project and these criteria have since been embodied in sequence prediction algorithms to identify CGIs (Ioshikhes and Zhang, 2000; Lander et al., 2001; Lee et al., 1998; Ponger and Mouchiroud, 2002; Waterston et al., 2002). However, the wealth of data has also highlighted the extraneous inclusion of non CGI sequences due to false discovery. Artifacts of sequence prediction are inevitable as they rely on arbitrary thresholds which are prone to oversimplifying biological data. The major cause of this overestimation comes from the incorrect identification of CpG and G+C rich repetitive elements as CGIs. As such, predicted CGI sets often contain spurious DNA sequences such as young Alu elements (Hellmann-Blumberg et al., 1993; Larsen et al., 1992; Rubin et al., 1994; Schmid, 1996; Takai and Jones, 2002). To combat this problem, algorithm prediction methods have been refined in two ways.

The first refinement investigated the threshold values for CGI length, CpG density and G+C composition. Contiguous 200bp sequences, identified by a sliding widow algorithm, were combined to generate entire predicted CGIs. Applying this method to a range of threshold values indicated that increased stringency reduced the identification of repetitive elements. A minimum length, CpG[o/e] and G+C composition of 500bp, 0.65 and 55% respectively,

significantly reduced the number of contaminating Alu elements identified by the algorithm (Takai and Jones, 2002). The increased stringency saw a reduction in the number of CGIs by approximately 92%. However, a reduced number of gene promoter associated islands, suggesting that bona fide CGIs were inadvertently being discarded.

The second refinement addressed the problem more directly, by a process called repeat masking. Repetitive elements and low complexity sequences are removed by screening against a comprehensive database of such elements (Repbase; (Jurka et al., 2005)). Preliminary analysis of the human genome sequence indicated that there were 50,267 islands, of which only 28,890 were identified as being unique (Lander et al., 2001). Masking is subject to iterative improvements due to the periodic increase of the Repbase repertoire (Jurka et al., 2005). Indeed reanalyzing the human CGI complement saw a further reduction of 1,890 CGIs, to give a more conservative estimate of 27,000 (Waterston et al., 2002). Clarification by repeat masking can be illustrated for a low copy repetitive element related to the adenovirus sequence. Located on Chromosomes 4 and 19[x], these elements are identified as single or tandem repeated CGIs by the ENSEMBL and NCBI prediction algorithms (Epstein et al., 1987). These sequences are recognized by repeat masker and eliminated.

Table 3.1-1. Details of CGI prediction algorithms

Database / Prediction	Length	G+C	CpG [o/e]	RM[a]	Comments	Ref
ENSEMBL	≥400	≥50%	≥0.6	N	stringent length constraint	(Birney et al., 2004)
NCBI relaxed	≥200	≥50%	≥0.6	N	Total CGIs = 307,193	
NCBI strict	≥500	≥50%	≥0.6	N	Total CGIs = 24,163	
USCS[b]	>200	≥50%	>0.6	Y	Total CGIs = 28,226	(Karolchik et al., 2008)
EMBOSS	UD[c]	UD	UD	NA	Variable parameters	(Rice et al., 2000)
CpGProD	>500	>50%	>0.6	Y	Total CGIs = 76,793	(Ponger and Mouchiroud, 2002)
CpGcluster	NA	NA	NA	N	Clustering. Total = 197,727	(Hackenberg et al., 2006)

[a]RM = Repeat Masked. Y=Yes N=No, NA=Non Applicable.
[b]Parameters used for CGI identification for the ENCODE project although totals vary due to repeat masking differences between hg17 and hg18 builds (Birney et al., 2007).
[c]UD = User Defined.

All major genome browsers employ CGI prediction algorithms based on these parameters and utilize repeat masking as a means to reduce the false discovery rate. The absolute numbers of CGIs is highly variable depending on the exact thresholds applied (Table 3.1-1). NCBI Mapview employs two different permutations of these parameters to provide a relaxed

[x] Chr4: Location (bp) 132864773 to 132905671 and Chr19: Location (bp) 41448637 to 41492694 / 42451390 to 42487246. Location based on Ensembl v46 NCBI build 36.

and stringent identification of islands (Table 3.1-1). NCBI strict predicts 24,163 unique CGIs whereas there relaxed criteria identifies more than 307,000. Such variability arises from the following observations: 1) the use of arbitrary thresholds, 2) no account taken for the heterogeneity of CGIs (Fig. 3.1-1) and 3) all sequence based prediction methods necessarily ignore methylation status.

These limitations may be addressed in the future by the incorporation of multiple layers of information into prediction methods, including methylation status and epigenetic modifications. Studies have shown that CGIs generally incorporate active histone marks and this can distinguish them within their long range chromatin context (Roh et al., 2005; Tazi and Bird, 1990; Weber et al., 2007). Genome wide epigenetic information will allow the refinement of prediction methods such that arbitrary sequence based predictions are supported by contextual biological information (Bock et al., 2007).

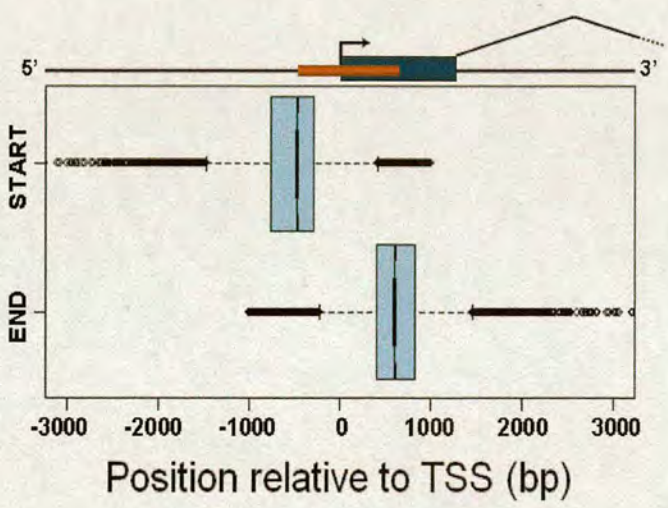


Figure 3.1-1. Schematic representation of an average CGI gene promoter.
The plot shows the distribution of start and end positions for all CGIs located within 1kb of a gene promoter (as annotated on the ENSEMBL genome browser). The average CGI (red bar; 1172bp) is indicated relative to the transcription start site (arrowed) and the 1st exon (dark blue bar). The boxplot indicates values for the median (central black line), 1st and 3rd quartiles (extremes of the blue box), maximum and minimum non outlying values (extremes of the dashed lines) and outliers (open circles). Data generate by Dr A. Kerr

3.1.3 Isolation of CpG Islands

As previously discussed, CGIs were initially identified through methyl sensitive restriction of genomic DNA. This procedure served as an analytical tool for the identification of CGIs but was necessarily destructive, and unsuitable for the isolation of intact CGIs. However an initial attempt to prepare and clone CGI sequences was through the isolation of larger digestion products (Shiraishi et al., 1995). Total genomic DNA was digested with a cocktail of restriction endonucleases which cut infrequently in CGIs[xi]. The digested DNA was then resolved by denaturing gradient gel electrophoresis. Fragments with a high GC composition

[xi] MseI (TTAA), Tsp5091 (AATT), Nla1 (CATG) and BfaI (CTAG).

were resolved whereas bulk genomic DNA was denatured and removed. Cloning and analysis of G+C rich fragments identified approximately 50% that were derived from CGIs. However the procedure was technically challenging and took no account of CpG clustering or lack of methylation, two important qualifiers for the identification of CGIs (Shiraishi et al., 1995).

Discovery of a nuclear protein with specific affinity for DNA sequences bearing methylated CpG dinucleotides facilitated an alternative means of CGI preparation (Lewis et al., 1992). Cross and Colleagues prepared a DNA affinity matrix by coupling a recombinant MBD from rat MeCP2 to a Nickel sepharose matrix (Cross et al., 1994). The MBD column was shown to selectively purify DNA fragments containing a high density of methylated CpG sites. Total human genomic DNA was digested with the restriction endonuclease MseI (TTAA). This had the effect of restricting bulk genomic DNA into small fragments (predicted average length of 123bp) while leaving CpG islands relatively intact (predicted average length of 625bp). This DNA was applied to the MBD column and eluted across an increasing NaCl gradient. Nonmethylated CGIs behaved as a poor ligand for the MBD, and eluted from the column at relatively low salt concentration. Methylated, CpG dense fragments were retained by the matrix and stripped from the DNA prior to further purification. Low affinity fragments including CGIs were fractionated a further twice prior to being fully methylated using the CpG methyltransferase, M.SssI. These were then reapplied to the column and fractionated. High salt fractions, containing the artificially methylated CGIs, were pooled and rechromtographed. High affinity DNA sequences were precipitated and cloned into pGEM 5zf- (Promega). Characterization of the cloned DNA fragments indicated that they were CpG and GC rich. Furthermore, they represented low copy number sequences with the exception of rDNA repeats (12.5%) which also resemble the sequence composition of CGIs (Bird et al., 1985; Bird and Taggart, 1980; Cross et al., 1994). It was confirmed that a significant fraction of the clones associated with the 5' ends of annotated genes. 80% of the library had significantly elevated CpG and G+C content suggesting that the cloned library represented a comprehensive CGI set.

This MBD technology has subsequently been applied to the purification of CGI from Mouse (Cross et al., 1997a), Chicken (McQueen et al., 1996) and Pig (McQueen et al., 1997). CGI libraries were also generated from human BACs, PACs and flow sorted chromosomes to allow detailed analysis of the relationship between CGIs and genes (Cross et al., 1999; Cross et al., 2000).

The original human CGI library was prepared, primarily, as a means of promoter identification. The purpose being to identify the transcription start sites and regulatory elements of cloned cDNAs prior to the availability of whole genome sequence. However, as the functional and regulatory significance of CGIs has become more apparent the utility of the library has increased. Novel estrogen responsive elements (EREs) were identified through screening the library with a recombinant Estrogen receptor (Watanabe et al., 1998). Subsequently, microarray probing technology has yet furthered the applications of the library. The arrayed CGI set has been used to screen CpG island methylation profiles in both normal and cancer tissues (Huang et al., 1999; Weber et al., 2005; Yan et al., 2001; Yan et al., 2002). Chromatin Immunoprecipitation (ChIP) has allowed the identification and characterization of transcription factor binding sites (Mao et al., 2003; Weinmann et al., 2002). To facilitate these applications, the entire library was re-sequenced and mapped to the human genome (Heisler et al., 2005). In total, 20,736 CGI clones were sequenced representing 9595 unique genomic loci. After justification and filtering the non-redundant proportion of the library represented 7184 unique CGI sequences, although this number contains some ribosomal repeats and non-specific contaminating DNA (Heisler et al., 2005).

The utility of the CGI library as a set of regulatory elements is clear. However represents less than 1/3rd of the human complement. This is partially due to the cumbersome nature of the purification process which resulted in the loss of some sequences. In addition, elevated GC density, a prerequisite for CGI sequence, is intrinsically difficult to sequence using conventional technology. As this library represents a biologically relevant set of regulatory elements, it would be of extreme value to generate a more comprehensive CGI set.

3.1.4 Aim

A limitation in many studies investigating the role of human CGIs has been uncertainty concerning their identification. As previously discussed, the criteria for designating a sequence as CGI-like are currently exclusively bioinformatic in nature, relying on the differences in the base composition and CpG frequencies (observed/expected) between bulk genomic DNA and CGIs. This obviates the need for a more biologically relevant method for CGI identification.

This chapter described the development and application of a novel chromatographic technique to fractionate CGIs and create a resource for future analysis. This technique,

termed CXXC Affinity Purification (CAP), selectively enriches for nonmethylated CpG rich sequences from bulk genomic DNA, and serves as an elegant preparative and analytical tool.

3.2 Results: CXXC Affinity Purification

The CXXC domain (PF02008) was first identified in the N-terminal region of the mammalian maintenance DNA methyltransferase, DNMT1 (Bestor, 1992). Removal of the N' terminal region of DNMT1, containing this motif, was shown to induce ectopic *de novo* methyltransferase activity (Bestor, 1992). Alignment has subsequently identified several additional mammalian proteins which carry homologous sequences. These include MLL (mixed lineage leukemia; also known as All-1 and HRX), MBD1 (Methyl CpG Binding protein 1, also known as PCM1), JHD1B (also known as Fbl10 or JEMMA) and CGBP (Fig. 3.2-1a, (Cross et al., 1997b; Koyama-Nasu et al., 2007; Ma et al., 1993; Voo et al., 2000)). BLASTp alignment to a protein database (refseq) indicates that nine human proteins have sequences which bear significant homology to the archetypal DNMT1 CXXC (Altschul et al., 1997)

The domain is characterized by a cysteine rich motif with the bipartite consensus sequence; $CX_2CX_2CX_nCX_2CX_2C$ (Bestor, 1992; Cross et al., 1997b; Ma et al., 1993). The two triplet cysteine repeats coordinate two Zn^{2+} ions to form a crescent like structure (Allen et al., 2006; Lee et al., 2001). NMR determination of the MLL CXXC showed that the secondary structure contains two short alpha-helices which form the basis of the zinc coordination module. All 6 cysteines are essential for DNA binding (Allen et al., 2006).

Alignment of homologous genes, show that the CXXC is highly conserved at the amino acid level. Paralagous CXXC3 domains of human and murine MBD1 show 100% sequence conservation. In addition orthologous CXXCs found in various mouse proteins display high levels of sequence conservation out with the core cysteine residues ((Carlone et al., 2002);Fig. 3.2-1b).

The majority of CXXC domains can bind DNA, and do so by interaction with a single nonmethylated CpG site. *In vitro* studies have indicated that CpG is both necessary and sufficient for binding and that this is ablated when the cytosine is methylated (Birke et al., 2002; Jorgensen et al., 2004; Lee et al., 2001). Evidence suggests, however, that CXXC can also bind to hemimethylated CpGs, although with significantly reduced affinity (Birke et al., 2002; Jorgensen et al., 2004). Despite the apparent DNA binding specificity of the CXXC

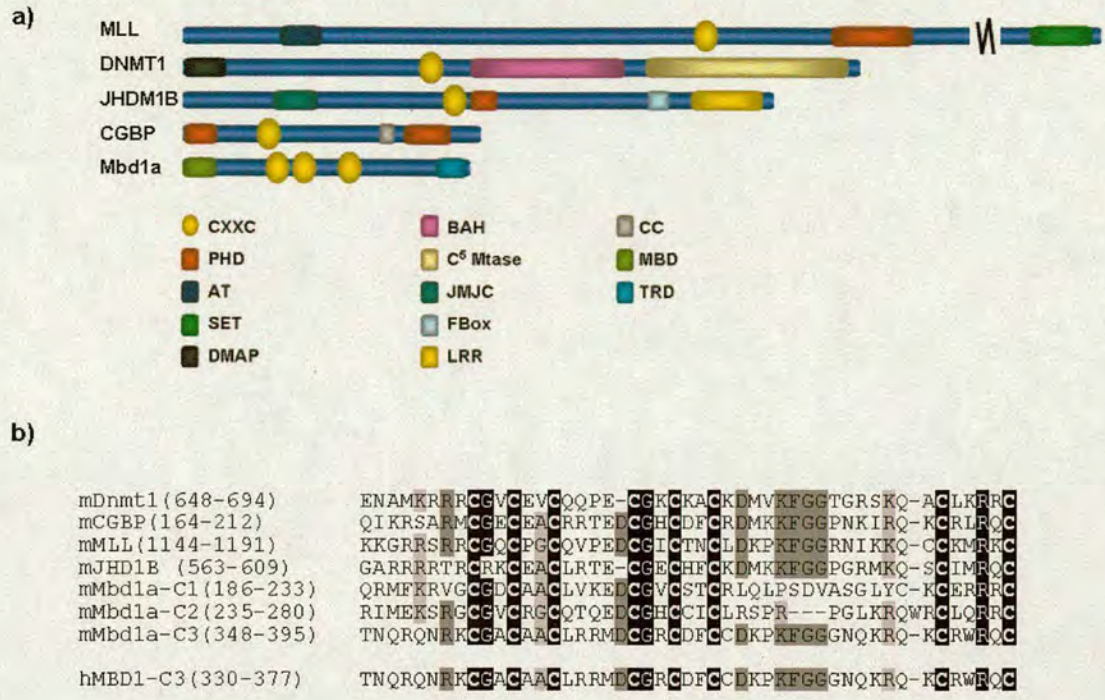


Figure 3.2-1. Conservation of the CXXC domain.
(a) Scale schematic representations of mammalian proteins bearing the conserved CXXC domain (filled gold ovals). Additional domains identified by the pfam protein database are illustrated and colour coded as per the legend. **(b)** Amino acid alignment of CXXC domains from various mammalian proteins. Conserved and partially conserved amino acids are shaded (black and grey respectively). The six essential core cysteine residues are emboldened.

domain, it is interesting to note that the CXXC domains 1 and 2 of mammalian MBD1 have no DNA binding affinity (Jorgensen et al., 2004). In addition, binding activity has not been reported for the DNMT1 CXXC and *In vitro* experiments suggest that it is non functional in this regard (unpublished observations).

The CXXC domain appears to act primarily as a DNA binding module. Mammalian MBD1 cannot localize to, or inhibit transcription from, nonmethylated promoters when lacking a functional CXXC3 domain (Fujita et al., 2000; Fujita et al., 1999; Jorgensen et al., 2004). CXXC Mutagenesis experiments showed complete loss of promoter transactivation in murine MLL (Ayton et al., 2004). In contrast however, DNMT1's CXXC does not bind DNA but is associated with a domain known to affect the specificity of the methyltransferase activity (Bestor, 1992). Mammalian MBD1 isoforms containing all 3 CXXC domains also seem to indicate a role outwith DNA binding. Two of the CXXC domains are incapable of DNA binding, but have been shown to interact with Ring1b, a component of PRC1. The third domain has been implicated in interaction with another PRC1 component; hPc2

(Sakamoto et al., 2007). Although data is limited, it appears that the CXXC motif may function as an important protein:protein interphase in addition to its well characterized role in nonmethylated DNA binding.

3.2.1 Recombinant CXXC expression and purification

The pET30-CxxC-3 expression plasmid was constructed by cloning the coding sequence of murine Mbd1a (amino acids 344 to 439) into the pet30b protein tagging vector (Novagen). The construct, encoding the entire third CXXC domain, was inserted into the EcoR1 and Hind111 sites of the multiple cloning site (MCS). The CXXC was cloned, in frame, with an N' terminal 6x histidine tag, thrombin cleavage consensus (endolytic serine protease), S tag and an enterokinase cleavage consensus (serine protease; Fig. 3.2-2, construct generated by Dr Helle Jorgenesen (Jorgensen et al., 2004)).

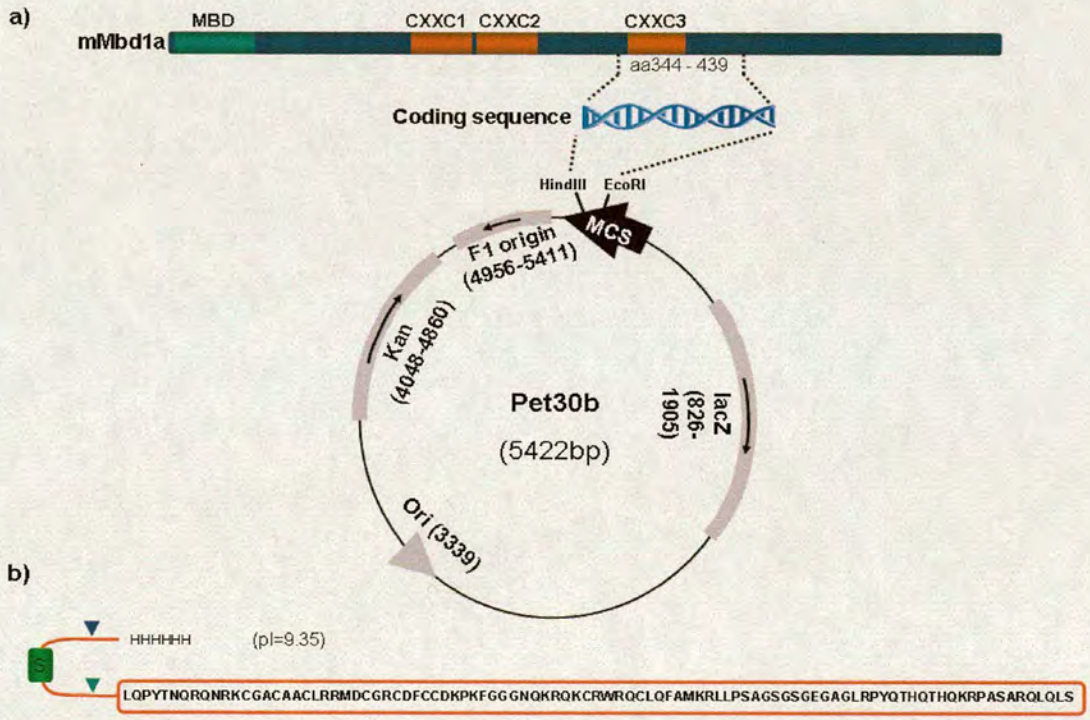


Figure 3.2-2. Cloning schema for the CXXC construct.
(a) Schematic representation of mMBd1a indicating the locations of the MBD (green) and CXXC (Red) domains. The location of the protein fragment cloned is indicated (aa344-439) and the DNA coding sequence corresponding to this site cloned into the Hind111 and EcorR1 sites of the pet30b protein tagging vector. The multiple cloning site (MCS, heavy black arrow), open reading frames (grey bars) and origin of replication (grey arrow) are indicated. **(b)** Sequence of the recombinant protein (red box). Eneterokinase cleavage site (green arrow head), S Tag (green box), thrombin cleavage site (blue arrow head) and the Histidine tag (6xH).

pET30-CxxC-3 was transformed into BL21(DE3)pLysS chemically competent *E.coli* cells, and the clonal cell culture was expanded to 8 liters. Recombinant CXXC expression was

induced by 3 hour incubation in the presence of 1mM IPTG at 37°C. CXXC induction was assessed by SDS PAGE (Fig. 3.2-3b).

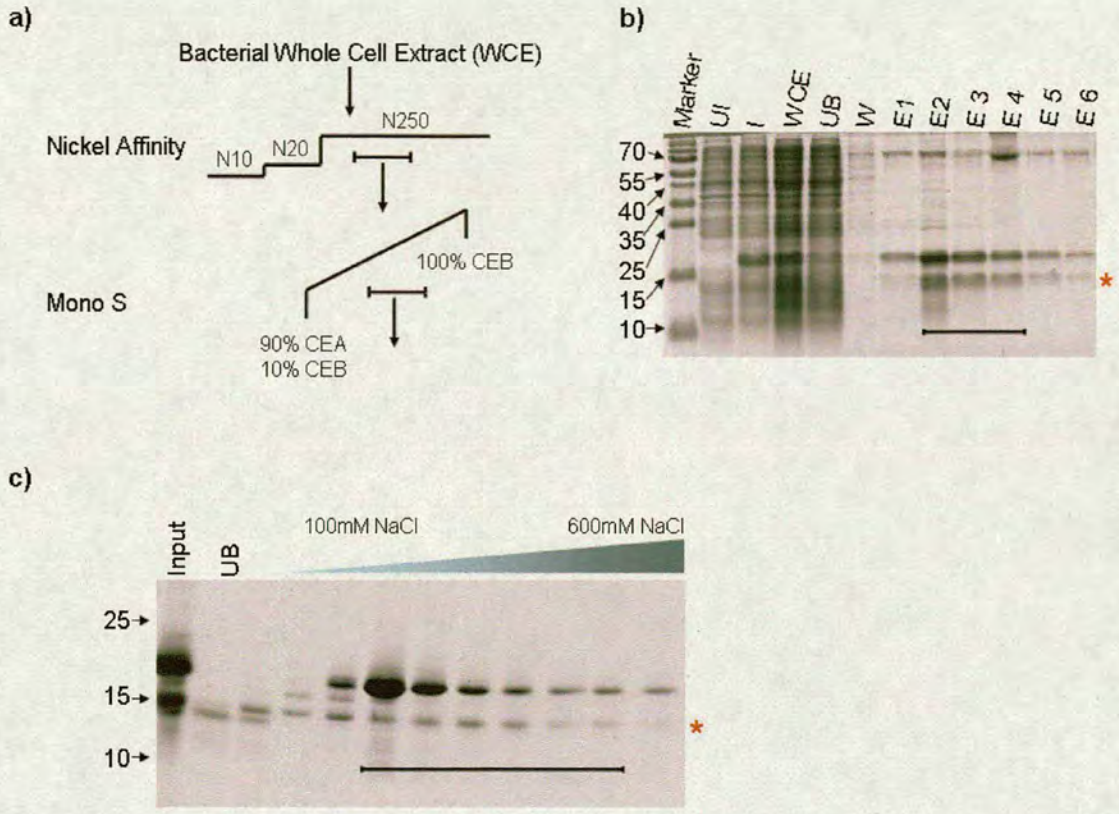


Figure 3.2-3. Purification of the recombinant CXXC3 domain of mMbd1a.
(a) Recombinant CXXC was purified from bacterial whole cell extract (WCE) by using the indicated two-step purification scheme (Ni-Affinity and MonoS). Chromatography buffers used for Nickel Affinity (N10, N20 and N250) and Mono S (CEA and CEB) are indicated. Fractions pooled from each step are bracketed. **(b)** CXXC induction and Nickel affinity purification, indicating protein content of uninduced bacterial culture (UI), IPTG induced culture (I), WCE, unbound protein (UB), wash (W), elutions (E1-6) and broad range marker (Marker, sizes indicated; Fermentas). Fractions to be further purified (bracketed) and the CXXC degradation product (red asterisk) are indicated. **(c)** Mono S cation exchange purification. Partially purified CXXC was bound to a 1ml Mono S chromatography column and eluted over an increasing NaCl gradient. Samples of input, Unbound (UB), elutions (100-600mM gradient) and size markers (not shown; indicated by labeled arrows) were resolved by SDS PAGE. Pooled fractions and the CXXC degradation product are indicated as for **(b)**.

Whole bacterial cell extract was prepared and bound to Nickel charged sepharose at 4°C in the presence of 10mM imidazole[xii] for 2hours. The CXXC adsorbed beads were washed in low Imidazole concentration buffer and then eluted across 8 successive fractions by

[xii] 6x polyhistidine binds strongly to divalent nickel ions by coordination between the imidazole group and the Ni^{2+} ions. Binding in the presence of low concentration Imidazole reduces non-specific binding of bacterial proteins with an inherent affinity for divalent cations. Proteins bound with high affinity can be eluted by competition with high concentration Imidazole or a reduction in buffer pH.

competition with high Imidazole concentration. SDS PAGE analysis of the fractions indicated that the nickel affinity purification had highly enriched a protein corresponding to the expected size of the recombinant CXXC (16.4kDa; Fig. 3.2-3b). However, contaminating proteins were still evident in the preparation so additional purification was required.

The basic nature of the CXXC construct (pI of 9.35) meant that it would be ideally suited to purification by cation exchange chromatography. Fractions containing large quantities of the purified CXXC (Fig. 3.2-3a & b; bracketed) were pooled and dialysed into 100mM NaCl cation exchange buffer. These fractions were bound to a 1ml Mono-S[xiii] cation exchange column (GE Healthcare) and fractionated across a 0.1-1M NaCl gradient. CXXC purity was assessed by SDS PAGE (Fig. 3.2-3c). At this stage recombinant CXXC was deemed pure enough for application to DNA chromatography. It is worth noting that recombinant CXXC always co-purified with a 2nd protein fragment presumed to be a degradation product (Fig. 3.2-3b and c - asterisk). Additional attempts to remove this degradation product were unsuccessful or resulted in unacceptable loss of total CXXC[xiv].

3.2.2 Activity and affinity of recombinant CXXC

The CXXC3 domain of mMb1a has been shown to bind to a single nonmethylated CpG site which is both necessary and sufficient for DNA binding. It has also been shown that binding affinity is lost when the cytosine base of the CpG is methylated (Jorgensen et al., 2004). To confirm the activity and specificity of the purified CXXC construct, DNA binding was characterized by bandshift assay.

A titration of recombinant CXXC was incubated with a methylated or nonmethylated, end labeled DNA probe containing 27 CpG sites. Binding assay reactions were resolved by agarose gel electrophoresis and imaged by phosphorImager. Increasing amounts of CXXC was shown to induce a gel mobility shift when incubated with the nonmethylated form of the probe but not when incubated with the methylated probe (Fig. 3.2-4). This confirmed that the recombinant CXXC had non-methyl specific binding with no detectable affinity for DNA containing only methylated CpGs. It is interesting to note that the protein:probe complex

---

[xiii] Mono S chromatography media is composed of the methyl sulfonate group (strong cation exchanger) conjugated to a highly crosslinked polystyrene matrix.

[xiv] To reduce the degradation observed the CXXC construct was prepared such that all the media contained Zn^{2+} ions and full protease inhibitors as this had previously been shown to stabilize constructs coordinated by metal ions. This however had little effect on the appearance of the degradation product Kelly, S.M., Pabit, S.A., Kitchen, C.M., Guo, P., Marfatia, K.A., Murphy, T.J., Corbett, A.H., and Berland, K.M. (2007). Recognition of polyadenosine RNA by zinc finger proteins. Proc Natl Acad Sci U S A *104*, 12306-12311..

formed, is not a discrete band suggesting that multiple CXXC domains can associate with each 135bp probe. This is an important observation and will be discussed later.

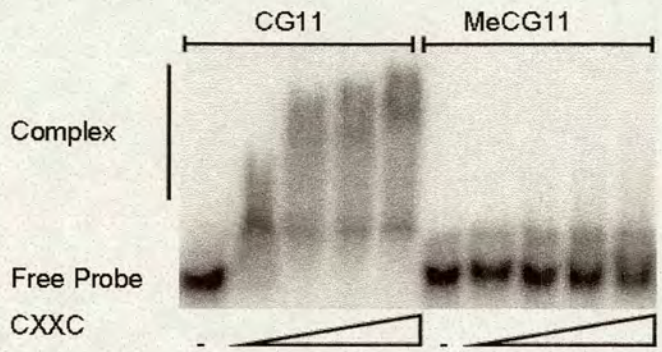


Figure 3.2-4. Purified CXXC specifically binds nonmethylated DNA.
Bandshift assay showing the CXXC complexed with a DNA probe containing 27 non-methylated CpG sites. Non-methylated probe DNA (CG11) or methylated probe (MeCG11) were incubated with 0, 250, 500, 1000 or 2000ng of recombinant CXXC protein.

3.2.3 Preparation of a CXXC chromatography column

The purified CXXC construct was dialysed into a sodium phosphate buffer[xv] and coupled to 1ml of nickel sepharose. Following removal of unbound protein, the beads were washed in a low concentration of imidazole (10mM) to remove weak affinity contaminants. The beads were then re-equilibrated to remove the imidazole and packed onto a 1ml chromatography column (Tricorn 5/50; GE Healthcare). The column was saturated with CXXC, containing approximately 45mg of CXXC, as determined by measurement of protein concentration in the input and unbound fractions. Prior to calibration, the CXXC column was equilibrated with a 0.1-1M NaCl gradient at 1ml/min. This pre-equilibration step ensured that the beads were thoroughly packed and that any residual bacterial DNA was eluted from the column (data not shown).

3.2.4 Plasmid fragment calibration

The resolution of affinity column chromatography techniques can be affected by a range of factors including matrix bed volume, the nature of the ligand and the amount of the affinity construct. As such it was necessary to calibrate the CXXC column in order to determine its efficacy for DNA chromatography. In order to do this, DNA fragments of varying length, CpG density and CpG methylation status were resolved across an increasing NaCl gradient.

[xv] HEPES, the buffering component for cation exchange, has an amine group which can reduce nickel resin.

The pABS plasmid, containing the CpG island for the mouse APRT gene, was linearised and restricted into 3 fragments using the endonucleases, EcoR1 and Hind111 (Macleod et al., 1994). The digestion products generated were 1.4, 2 and 2.7kb in length with CpG densities[xvi] of 0.22, 0.5 and 1.0 respectively (Fig. 3.2-5a).

1μg of the fragmented plasmid was treated with the CpG methylase M.Sss1 and 500ng aliquots of the methylated and nonmethylated DNA were end labeled with α^{32}P-dCTP. The CXXC chromatography column was attached to an FPLC system (Pharmacia) and pre-equilibrated with 100mM NaCl running buffer. The nonmethylated fragments were bound to the column in 100mM NaCl buffer and then eluted across a 30ml, 0.1-1M salt gradient (Fig. 3.2-5b). This process was repeated for the methylated fragments, and 1/3rd of each fraction was ethanol precipitated.

A relative measure of DNA content in each fraction was obtained by measuring the incorporated label using a Tri-Carb 2100TR liquid scintillation counter (Perkin Elmer). Fractions were then resolved by agarose gel electrophoresis to determine the specific elution profile of the methylated and non-methylated fragments across the NaCl gradient.

All three methylated fragments eluted at 400mM NaCl concentration irrespective of CpG density. At this low NaCl concentration the CXXC matrix showed retention of all non-methylated DNA fragments. Unlike the methylated DNA fragments however, increasing the NaCl concentration resolved nonmethylated DNA as a function of CpG density. The CpG deficient fragment ($CpG_{[o/e]}$ of 0.22) eluted at 550 to 600mM NaCl whereas 800-900mM NaCl was required for the elution of the CpG dense fragments (Fig. 3.2-5b). When methylated and nonmethylated fragments were pooled and fractionated the CXXC column could selectively resolve the nonmethylated CpG rich DNA fragments from CpG deficient or methylated DNA fragments (data not shown).

The DNA binding calibration indicated that all DNA fragments bound to the CXXC affinity matrix but that methylated DNA did so with reduced affinity. Nonmethylated DNA retention was shown to be CpG density dependent with increased CpG content binding at a higher NaCl concentration. This suggested that the CXXC matrix formed multiple contacts with

[xvi] CpG density was determined by calculation of the CpG[observed/expected] ratio. Expected values were calculated using the equation: Expected = ((number of cytosines/fragment length)*(number of guanines/fragment length))*fragment length.

each DNA fragment and that CpG dense fragments represented higher affinity ligands for the affinity column. This was consistent with the observation that the bandshift probes were complexed with multiple CXXC domains. This result confirmed that CAP could selectively enrich for nonmethylated sequences from a low complexity pool of CpG deficient and methylated DNA fragments. However, to determine if CAP could effectively purify CGIs from total genomic DNA further calibration was necessary.

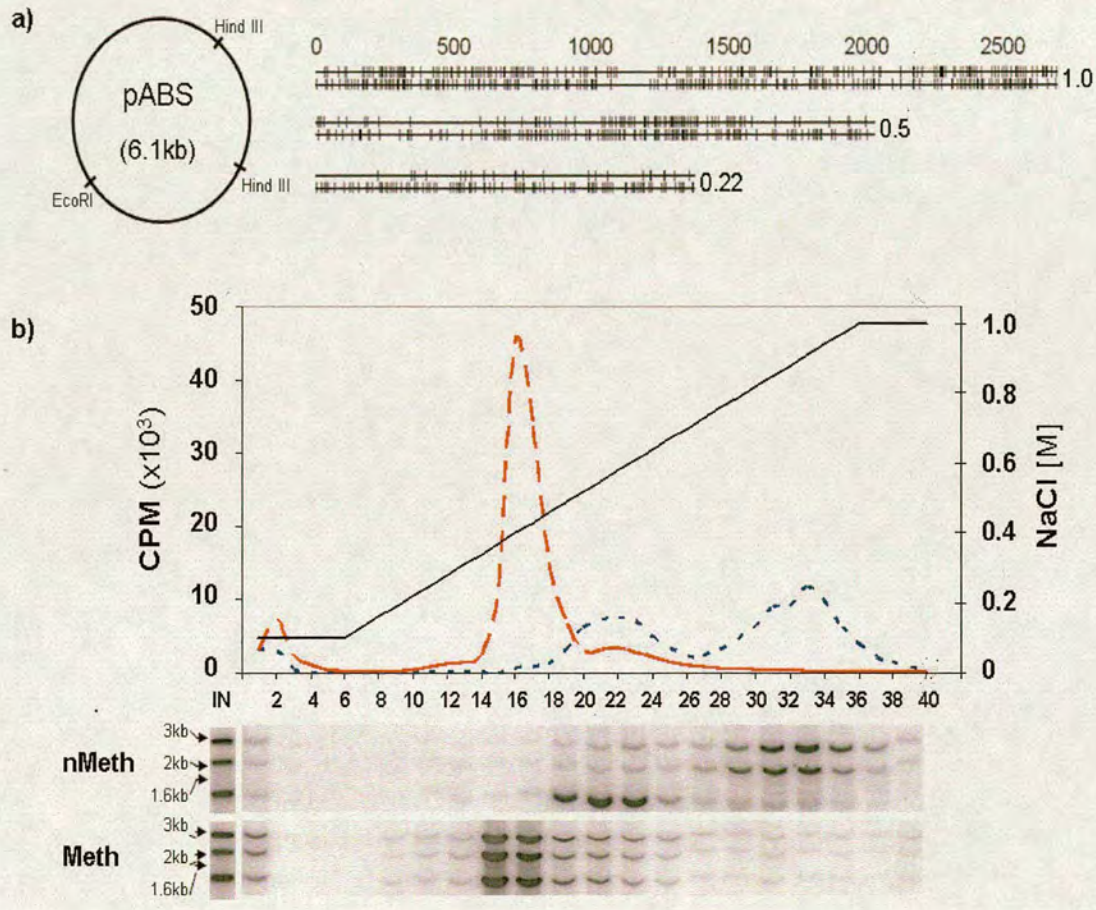


Figure 3.2-5. Nonmethylated DNA binding is dependent on CpG density.
(a) Schematic representation of the pABS plasmid and three restriction products. The CpG [o/e] values are indicated (at right of fragments), CpG (top strand) and GpC (lower strand) are indicated by black strokes. The length of each fragment is indicated by the scale (above the upper most fragments, base pairs (bp). (b) The upper panel indicates the elution profile of methylated (dashed red line) and nonmethylated (dashed blue line) fragments across the NaCl gradient (solid black line). All methylated fragments elute at low NaCl concentration (~0.4M) whereas nonmethylated fragments don't elute until (0.6M) with the most CpG dense fragments being retained by the CXXC matrix until (0.9M). Size markers, as determined by the 1kb+ DNA ladder (Invitrogen, not shown) are indicated (black arrows and corresponding sizes). CPM (counts per minute) represents an arbitrary measure of radiolabeled DNA.

3.2.5 CAP Genomic DNA calibration

To determine if CAP could specifically enrich for nonmethylated CGI sequences, total human genomic DNA (male and female) was restricted with the endonucleases MseI (TTAA) and fractionated over the CXXC column. The rationale behind this digestion has previously been discussed (Cross et al., 1994). Mse1 restriction was utilized as it cuts AT-rich genomic DNA into small fragments, whilst leaving CGIs relatively intact. Bulk genomic DNA has a CpG on average every 100 base pairs so the majority of Mse1 fragments have insufficient CpGs to be retained by the CXXC matrix. CGIs on the other hand, with one CpG per approximately 10 base pairs, will give rise to long CpG rich fragments with high affinity for the CXXC column.

The NaCl gradient was optimized to incorporate a 600mM NaCl wash to increase the separation between the methylated and non-methylated DNA fragments (Fig. 3.2-6a). Eluted fractions were interrogated by PCR using primers specific for a range of known CGIs and non-CGI sequences (Fig. 3.2-6b). The non-methylated CGI of the *P48* gene eluted at high salt. The X-linked monoamine oxidase (*MAO*) gene eluted as a single high salt peak from male genomic DNA (where it is nonmethylated), but as two separate peaks at low and high salt when female DNA (with one methylated and one nonmethylated allele) was fractionated. The CGI associated with the testis specific antigen gene *NYESO* (somatically methylated) eluted from the CXXC column at low salt as predicted (Fig. 3.2-6c; (De Smet et al., 1999)). The data confirms that CAP can effectively purify nonmethylated CGIs from total genomic DNA.

The CXXC column was used infrequently, and as such it was difficult to determine the lifespan of the matrix. DNA fractionations were consistent across a six month period; although after 12 months the column no longer binds to methylated DNA (data not shown).

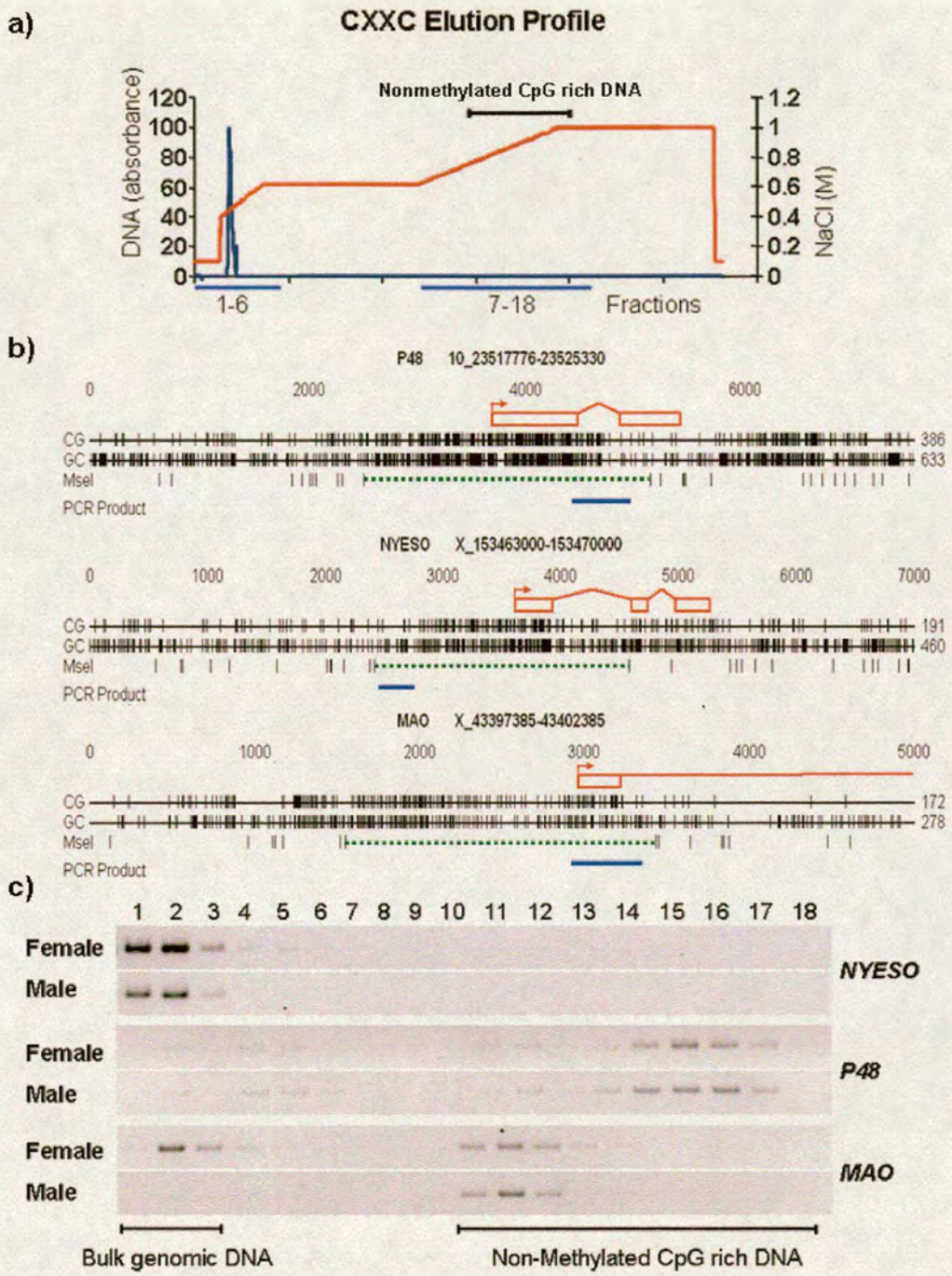


Figure 3.2-6. CAP enrichment of nonmethylated CGI sequences

(**a**) A typical elution profile of bulk genomic DNA (blue line) from a CXXC affinity chromatography column. Genomic DNA (100µg) was applied to the CXXC affinity matrix (see Methods) in low salt (0.1M NaCl) and eluted with a gradient of increasing NaCl (red line). Eighteen fractions were interrogated by PCR (blue lines). The bracket above indicates fractions that were found to contain non-methylated CGIs. **(b)** Sequence characteristics of specific CGIs of known methylation status (*NYESO, P48* and *MAO*). The direction of transcription (red box) is arrowed. MseI fragment assessed for column binding (dashed green line) and the region amplified by PCR (blue line) are indicated. (**c**) Elution of specific CGI sequences of known methylation status. Methylated CGIs (*NYESO* and *MAO* in females) coelute with bulk genomic DNA (see bracket and equivalent fractions in (a) whereas non-methylated CGIs (*P48* and *MAO*) elute at high NaCl concentration.

3.3 Results: Generation of a novel CGI set

3.3.1 CXXC affinity purification of a human blood CGI fraction

Having demonstrated the functionality of the CAP procedure it was applied to the purification of a complete CGI library. To this end, we prepared whole blood DNA from three healthy male donors. Blood was used in preference to sperm, as it is known that CpG rich repeats and GC rich telomere proximal regions are unmethylated in sperm and would represent an unacceptable contamination of the CGI fraction (Brock et al., 1999; Kochanek et al., 1993). Genomic DNA was MseI digested and applied to the CXXC column. The DNA was eluted across an increasing NaCl gradient and the high salt fractions (0.8-1M) pooled and re-chromatographed for a second time. This additional wash was introduced to ensure efficient removal of bulk genomic DNA contaminants from the CGI fractions. CGIs retention was confirmed by PCR validation of the first and second column runs. Pooled CGI fractions were precipitated and cloned into the NdeI site of the pGEM5zf- (Promega) plasmid vector to generate a CGI library.

A panel of 181 cloned MseI fragments was sequenced using the Sanger dideoxy method (Sanger et al., 1977). The sequences were aligned against the human genome to determine the origin of the cloned MseI fragments. 145 of the MseI inserts were derived from human genomic sequence, 94% (136) of which showed elevated CpG[o/e] and/or G+C composition relative to the genome average (Fig. 3.3-1a). Full details of the alignment are summarized in Table 3.3-1. The insert length were also increased relative to the expected MseI fragment size for bulk genomic DNA (123bp; Fig. 3.3-1b dashed line). The sequence composition of the clones was highly suggestive of a library enriched for CGI fragments. Furthermore, none of the cloned sequences showed redundancy (with the exception of rDNA repeats), which suggests comprehensive library coverage.

Table 3.3-1. Pilot sequence characteristics of the CGI library.

Sequence characteristic	Number of Inserts
Failed Sequence	19
Cloning vector sequence	8
E. coli genomic sequence	9
Human: Non CGI sequence	9
Human: CGI sequence	126
Human: rDNA sequence	10

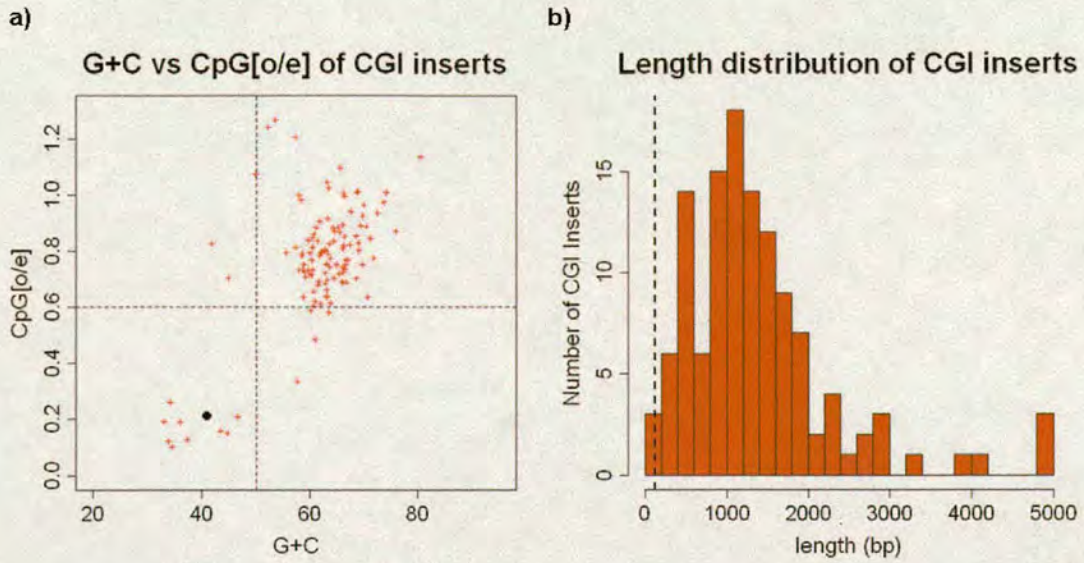


Figure 3.3-1. CAP enriches for sequences with classical CGI sequence properties. **(a)** A panel of CAP purified CGI inserts show elevated CpG and G+C content compared to bulk genomic DNA (black dot). The majority of inserts fulfill the minimum sequence criteria of ≥ 50% G+C (vertical dashed line and CpG [o/e] of ≥ 0.6 (horizontal dashed line) commonly used to identify CGI sequences (Gardiner-Garden and Frommer, 1987). **(b)** Cloned inserts are longer than that expected for bulk genomic DNA (123bp, dashed line) consistent with reduced TTAA recognition sites in G+C rich CGIs.

3.3.2 Sequencing the CGI library

The optimized library[xvii] was sequenced to redundancy by the sequencing facility at the Welcome Trust Sanger Institute. In total approximately 172,000 colonies were sequenced with paired end reads, representing 119,487 genomic templates. Sequences which mapped to unique genomic locations[xviii] were aligned to 28,013 unique MseI fragments. Plots depicting insert length versus either G+C content or CpG[o/e] indicated elevated CpG and G+C composition of the cloned fragments (Fig. 3.3-2a,b). Despite these general characteristics, a population of smaller inserts resembled the sequence characteristics of bulk genomic DNA (Fig. 3.3-2a,b dashed line). To determine if these short fragments represented library contamination or genuine CGI sequences, a panel of inserts (<512bp) were tested for CXXC binding. Human male DNA was MseI digested and fractionated using the CXXC column. Primers, complementary to short MseI fragments, were used to identify these sequences by PCR amplification from input and CXXC bound DNA (Fig.. 3.3-2c). Each MseI fragment was then allotted to a quadrant determined by a G+C vs. CpG plot (Fig. 3.3-2d). Only fragments satisfying both elevated G+C (>50%) and CpG[o/e] (>0.6) criteria were retained by CAP and represented genuine nonmethylated CGI inserts. As such, MseI fragments

[xvii] An efficiency of >2million colonies per pGEM ligation was required for library coverage.
[xviii] Human genome assembly NCBI36.

shorter than 512bp with a G+C content of <50% and/or a CpG[p/e] of <0.6 were filtered out of the CGI set (filtered islands are represented as grey dots in Fig. 3.3-2a,b). Remaining sequences were compiled into complete CGIs by knitting together contiguous fragments[xix].

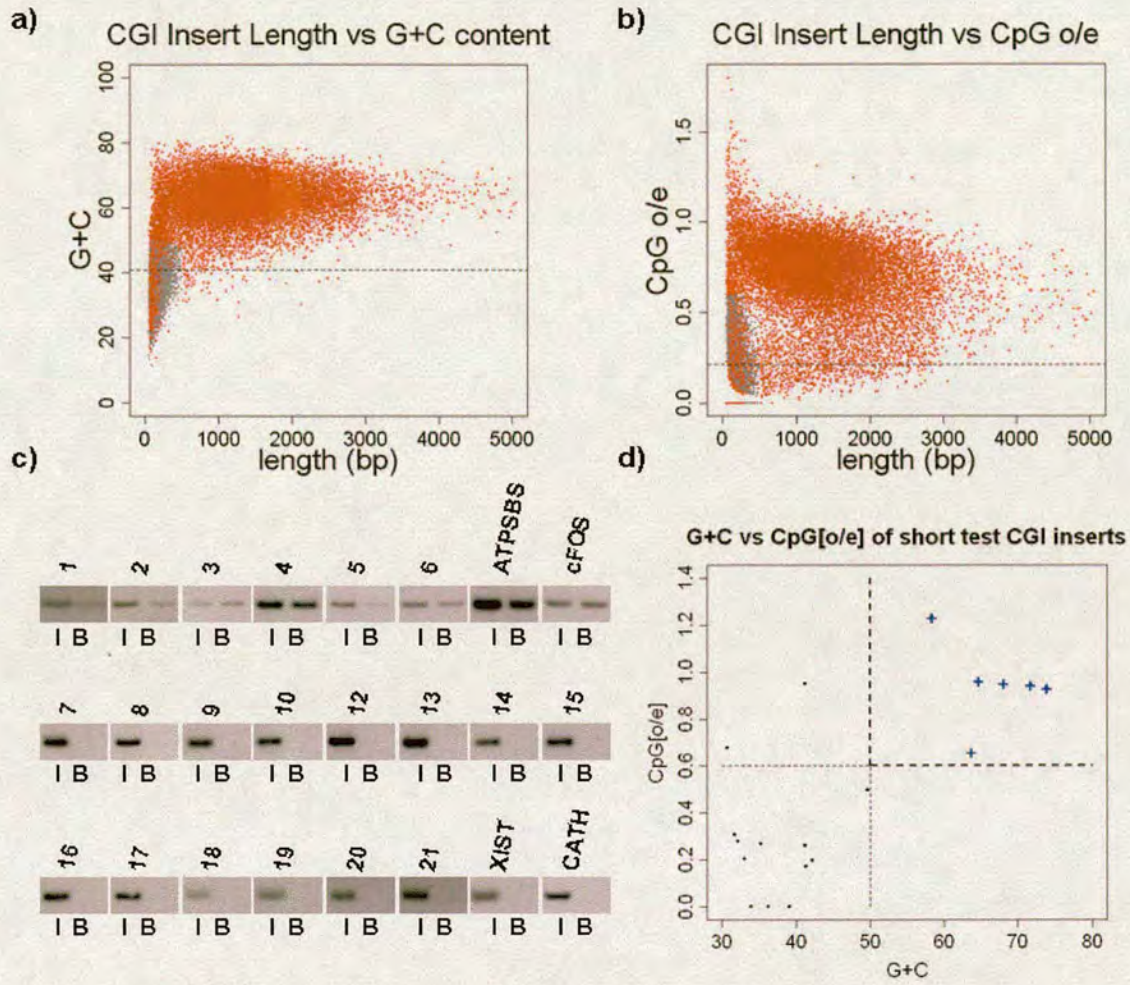


Figure 3.3-2. CGI sequences characteristics and CAP binding
Plots of fragment length versus G+C content (**a**) and CpG[o/e] (**b**) for 28,013 unique Mse1 inserts. Fragments shorter than 512bp with a G+C content = <50% and a CpG[o/e] = <0.6 (grey dots) are indicated. The dashed line indicates the base composition **(a)** and CpG o/e **(b)** of bulk genomic DNA. **(c)** Investigation of CAP binding to short DNA inserts. PCR interrogation of short CGI library sequences from input (I) and CAP retained (B) DNA. Regions amplified are indicated by numbers or, in the case of positive (*ATPSBS* and *cFOS*) and negative (*XIST* and *CATHEPSIN*) controls, as gene names. **(d)** Sequence characteristics of short inserts which bound (+) or did not bind (black points) the CXXC column.

Longer CGI inserts were not filtered from the set, as reduced CpG and G+C density for these fragments likely resulted from the cloning of fragmented CGIs. These sequences were retained due to their utility in localizing genuine CGIs. The final refined set contained 17,387 unique CGIs with sequence composition consistent with classical CGIs (Fig. 3.3-3).

[xix] Fragments were combined if contiguous fragments were separated by ≤ 100bp.

This set can be viewed via the 'CPG island' DAS track hosted by the ENSEMBL genome browser (http://www.ensembl.org/ Homo_sapiens/index.html; (Birney et al., 2004), Dataset 1). Elevated G+C content of the CGI set hindered DNA sequencing, such that only approximately 70% of clones were identified. This suggests that the total number of CGIs may be in excess of 25,200.

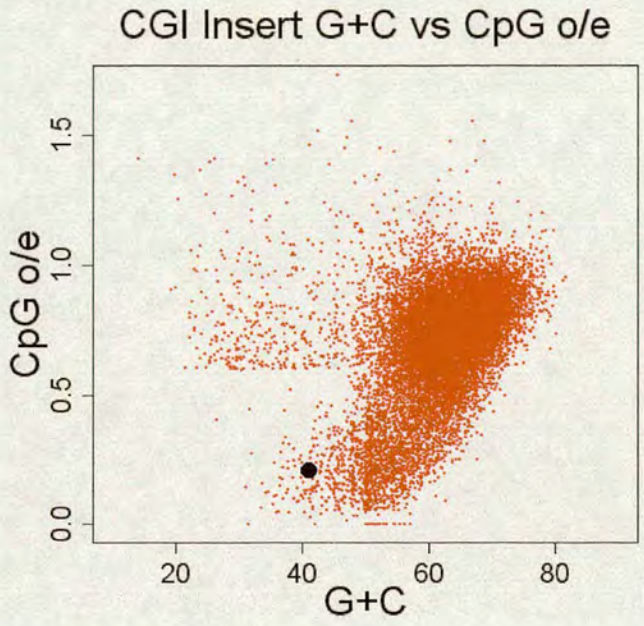


Figure 3.3-3. A CAP purified library represents a comprehensive CGI set.
A filtered insert set representing 17,387 CGIs, show a discrete distribution that is distant from bulk genomic DNA (black dot).

3.3.3 Distribution of the CAP CGI library

Studies have indicated that CGIs are not distributed evenly throughout the human genome (Craig and Bickmore, 1994; Lander et al., 2001). To investigate this further, all CGIs were plotted against the 24 human chromosomes to generate a physical map of CGI location (Fig. 3.3-4a). The schematic obviates gene rich chromosomes such as 19, which is overrepresented in the CGI set containing one CGI every 57kb. Conversely, gene deficiency on chromosomes 13 concords with a paucity of CGIs represented by an average density of one island every 303kb. Genome wide comparison identifies a striking correlation between CGI density and gene distribution (R=0.96; Pearson's correlation; Fig. 3.3-4b). Whereas, physical length of the chromosomes correlates far less with CGI density (R=0.58; Pearson's correlation; Fig. 3.3-4b). These findings are consistent with the co-localization of CGIs with gene promoters and regulatory elements.

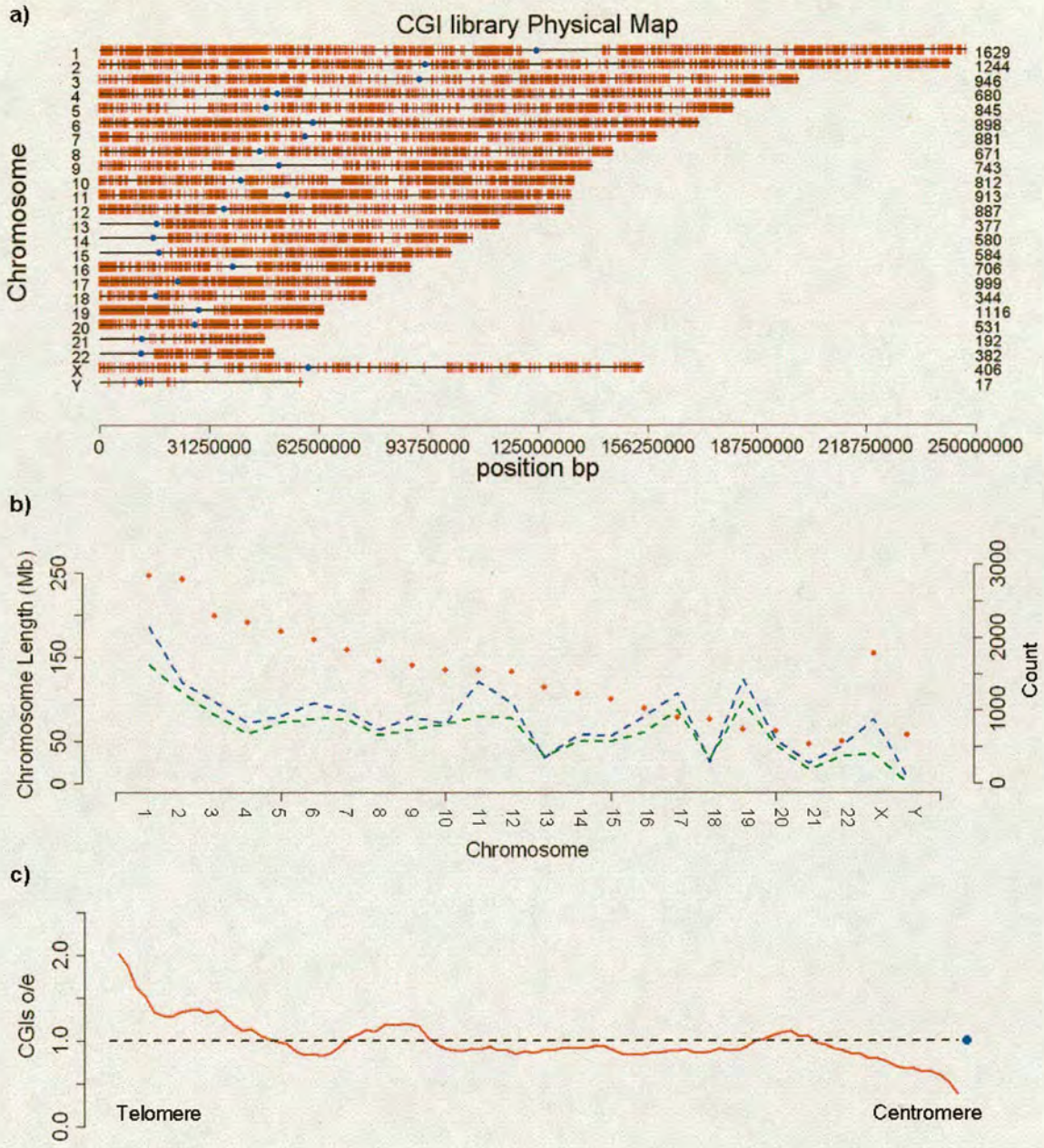


Figure 3.3-4. Genomic distribution of the CGI set.
(a) The distribution of cloned CGIs (red strokes) on human chromosomes. The number of CGIs on each chromosome is shown (right) and centromeres are denoted by blue dots. **(b)** A plot indicating the chromosomal length (red points), number of protein coding genes per chromosome (dashed blue line) and number of CGIs per chromosome (dashed green line) indicates a correlation between CGI and gene densities. **(c)** A plot depicting the average observed/expected CGI distribution (red line) along the autosomal arms indicates a relative enrichment proximal to the telomeres. The plot represents a window size and sliding resolution of 5% and 1% of autosomal arm length respectively. Values indicated are observed numbers of CGIs / that expected from even CGI distribution [o/e]. An [o/e] ratio of 1 (dashed line) and centromere position (blue point) are indicated.

Fluorescence in situ hybridization (FISH) experiments utilizing sequences enriched for CGIs, identified high density clustering of CGIs in T bands proximal to telomeres (Craig and

Bickmore, 1994). To determine if the CGI distribution indicates a similar trend and investigate spatial CGI density along the chromosome arms, an average schematic representation of chromosomal distribution was plotted (Fig. 3.3-4c). The representation of CGIs per chromosomal region was calculated using sliding windows of 5% with a periodicity of 1%. CGI distributions were calculated for all chromosomal arms and combined to give an average CGI density map. The observed number of CGIs was normalized against 5% of the total set to give a measure of density distribution. This plot indicated that while CGI levels approximate that expected along the majority of the chromosome, that CGIs are enriched in subtelomeric regions. Interestingly, the inverse is true for density surrounding the centromeres. This may be explained by the dearth of CGIs in pericentric heterochromatin such as illustrated for chromosomes 1 and 9 (Fig. 3.3-4a).

In addition to the genome wide distribution, it was of interest to investigate the specific association between CGIs and genes. In order to do this, library inserts were mapped relative to all genes annotated by the ENSEMBL genome browser. This showed that 76.3% of all islands associated with an annotated protein coding gene (see Materials and Methods for gene mapping). However it was interesting to note that only 49.5% of these mapped to the annotated TSSs of protein coding genes. This suggests that approximately half of all CGIs are located distal to promoters and as such may represent unannotated regulatory regions or alternative internal transcriptional initiation sites. Furthermore, it is likely that at least a proportion of the intergenic 'orphan' islands represent genes which are as yet unannotated, due to low level or highly restricted expression patterns. Gene overlap suggests that only 43.5% of protein coding genes have a CGI associated with their promoter (Table). However this likely represents the reduced coverage due to sequence failure. The discord between the observed and expected number of CGI gene promoters allowed a refinement of the estimated total. The fraction of genes observed to associate with a promoter CGI (43.5%) and the

Table 3.3-2. Relationship between CGI library inserts and genes.

Type of Overlap	Gene Biotype	Total Genes (CGIs)	Overlap with CGI (Gene)	Percentage	Overlap with CGIs (TSS)	Percentage
Gene Overlap: CGI	Protein	21384	15118	70.7	9312	43.5[a]
	All genes[b]	31524	15433	49	9529	30.2
CGI Overlap: Genes	Protein	17387	13271	76.3	8491	48.8
	All genes	17387	13360	76.8	8611	49.5

[a] The fraction of genes with promoter CGIs is less than the known fraction (56%) because of 31% sequence failure in the CGI set.
[b]All genes as classified on the ENSMBL genome browser including; non-coding RNAs, pseudogenes, VDJ regions etc.

fraction previously determined (56%) would suggest that there are 22,400 (Larsen et al., 1992). This figure is rather less than the 25,200 previously posited.

3.3.4 CGI prediction vs. the CAP CGI Set

CGI prediction algorithms identify CGIs based on three key parameters: CpG density, G+C content and sequence length. Most methods, employ a minimum CpG[o/e] density of 0.6 and minimum G+C content of 50%. However, the length of sequence which must satisfy these parameters varies between 200 and 500bp, the consequences of which are outlined in the introduction. Since, sequence based algorithms necessarily ignore methylation status it would be informative to compare bioinformatically identified islands with the more biologically relevant CAP CGI set. For this, two of the most commonly applied algorithms were compared with the library. The NCBI genome browser suite maintains two such algorithms with minimum length requirements of 200bp and 500bp which we formalize as $NCBI^{relaxed}$ and $NCBI^{strict}$ respectively. The total number of CGIs is summarized in table 3.1-1. 13,305 (76.5%) CAP CGIs are identified by $NCBI^{strict}$, which represents only 56% of all of the 24,163 islands predicted by this algorithm. Improved representation is provided by $NCBI^{relaxed}$, with 15,799 (90%) of the CAP CGI library being predicted. However despite the concordance between the two methods, the CGI library only represents 5.2% of the total 307,193 predicted. Additionally the overlap between the two prediction methods is minimal (7.9%) suggesting that the majority of CGIs detected are false positives. This relatively poor representation of the CGI library in the $NCBI^{strict}$ set is likely due to library sequence failure. However the apparent accordance between predicted numbers of CGIs by these two methods masks a significant difference. The CGI library identifies approximately 4,082 CGIs (23%) missed by the algorithm.

Table 3.3-3. Gene association: CpG islands missed by NCBI strict

Gene overlap	Number of CGIs	Percentage of CGIs
5'	778	19.1
3'	176	4.3
Intragenic	1421	34.8
Intergenic	1707	41.8
Total	*4082*	*100*

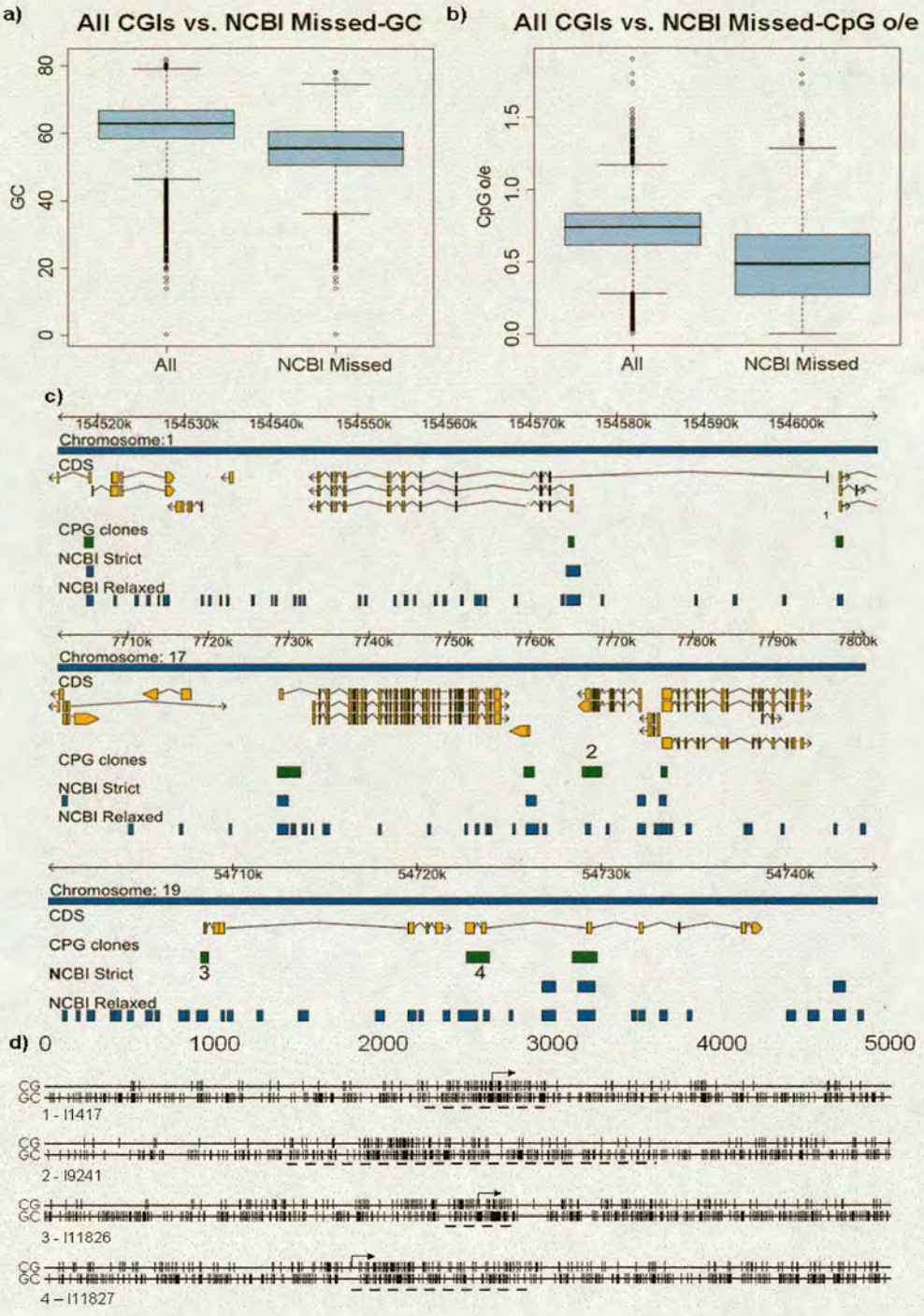


Figure 3.3-5. Sequence characteristics of CGIs missed by NCBIstrict.
Boxplots of G+C **(a)** and CpG[o/e] **(b)** indicate that islands retained by CAP but missed by NCBI strict have significantly reduced G+C base composition (p.value<2.2e-16) and CpG[o/e] (p.value<2.2e-16; NCBI-missed and all CGIs n=4082 and 13305 respectively). Boxplots are represented as for Fig. 3.1-1. **(c)** Three random chromosomal regions showing CGI sequences mapped to the human genome (green bars). Also shown, are CGIs predicted by the NCBI^{strict} and NCBI^{relaxed} algorithms (blue bars). The direction of transcription of coding sequences (yellow bars) is arrowed. Numbered CGIs (1-4) represent sequences not detected by the NCBI strict algorithm. **(d)** CpG maps of the four CGI clones not predicted by NCBI^{strict}. Transcription start sites in examples 1, 3 and 4 are indicated by arrows. Sequenced MseI fragments (dashed lines) and CpG and GpC sites (black strokes) are indicated.

Islands not predicted by this algorithm were, as predicted, weaker with respect to CpG and G+C content than the whole CGI set (Fig. 3.3-5a and b). Therefore, to ensure that these were not representative of genomic contamination, gene association was compared with the whole CGI set. Of four randomly selected islands represented in this category three were found to colocalise with the promoter regions of protein-coding genes (Fig. 3.3-5c,d). All CGIs (n=4082) retained by CAP but not predicted by NCBI strict were mapped relative to annotated human genes[xx]. The majority of non predicted islands are gene associated, although with an elevated level of intragenic islands (Table 3.3-3). These findings are consistent with the previous observation that intragenic islands are generally weaker than those associated with gene promoter regions (Ioshikhes and Zhang, 2000). These findings suggest that the 23% of cloned but not predicted islands represent genuine CGIs and that this method allows the identification of weaker islands by their lack of methylation. Based on this it is unlikely that sequence based algorithms can be significantly refined to include these CGIs without introducing large numbers of false positives as illustrated by NCBI^{relaxed}.

3.4 Discussion

Previous attempts to purify CGIs from genomic DNA were inefficient or cumbersome in their application (Cross et al., 1994; Shiraishi et al., 1995). The most comprehensive set of human CGIs was prepared using a reverse chromatographic procedure employing the MBD domain of rat MeCP2 (Cross et al., 1994). Whilst proving effective, technical limitations rendered the resulting library incomplete (Cross et al., 1994; Heisler et al., 2005). This chapter describes the development of a novel, preparative and analytical procedure to resolve these issues. A recombinant CXXC domain, with *In vivo* and *In vitro* specificity for non methylated CpG sites was purified and adsorbed onto a nickel sepharose support (Birke et al., 2002; Jorgensen et al., 2004; Lee et al., 2001). DNA Chromatography conditions were optimized to ensure the efficient removal of methylated and CpG deficient sequences. The rusulting affinity matrix permitted the isolation of CGIs from total genomic DNA (Fig. 3.4-1) in a single, nondestructive fractionation step.

3.4.1 Purifying and characterizing a Comprehensive CGI set

A comprehensive set of CGIs would not only further elucidate their role in the genome it would also provide an analytical tool representing the promoter and regulatory regions for the majority of human genes. Similar clone sets have been applied to the large-scale elucidation of DNA methylation profiles and identification of transcription factor binding

[xx] Genes were annotated according to ENSEMBL release 46 of NCBI36.

sites ((Huang et al., 1999; Mao et al., 2003; Watanabe et al., 1998; Weber et al., 2005; Weinmann et al., 2002; Yan et al., 2001; Yan et al., 2002)). Due to the lack of coverage provided by previous libraries a human somatic CpG island set was generated by CAP fractionation of male blood genomic DNA.

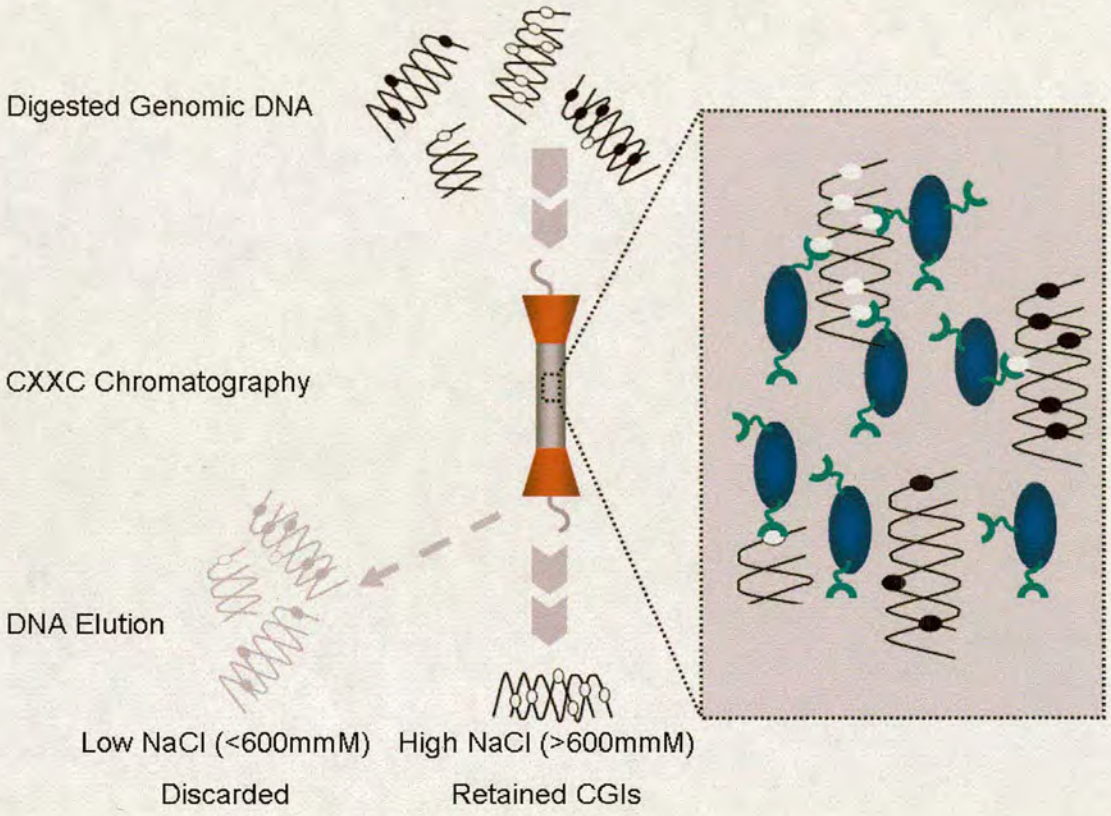


Figure 3.4-1. CXXC Affinity Purification (CAP).
Mse1 digested genomic DNA is bound to the CXXC affinity matrix in low NaCl and then eluted across an increasing salt gradient. Bulk genomic DNA containing dispersed methylated CpGs (filled circles), interacts minimally with the CXXC domains (green "Cs") and elutes at low NaCl and is discarded. CGIs containing clusters of nonmethylated CpGs (open circles) bind strongly to the CXXC matrix and are retained, allowing their purification at high NaCl concentration. The nickel sepharose support is represented by dark blue ovals.

Sperm DNA, known to have low levels of genomic methylation, was not the preferred material for the generation of this library as CpG rich repetitive elements, such as Alus[xxi],

[xxi] Alu elements are represented by approximately 1 million genomic copies in the haploid human genome Chesnokov, I.N., and Schmid, C.W. (1995). Specific Alu binding protein from human sperm chromatin prevents DNA methylation. J Biol Chem *270*, 18539-18542, Hellmann-Blumberg, U., Hintz, M.F., Gatewood, J.M., and Schmid, C.W. (1993). Developmental differences in methylation of human Alu repeats. Mol Cell Biol *13*, 4523-4530, Rubin, C.M., VandeVoort, C.A., Teplitz, R.L., and Schmid, C.W. (1994). Alu repeated DNAs are differentially methylated in primate germ cells. Nucleic Acids Res *22*, 5121-5127..

would bind tightly to the CXXC matrix. As such it was considered preferable to sacrifice a small proportion of relevant CGIs in favour of background reduction that these elements would otherwise represent (Rubin et al., 1994; Weber et al., 2007).

The CGI fraction was cloned and sequenced to redundancy. After filtering out false positives, the clone set had sequence characteristics which epitomized classical CGIs. The library, consisting of 17,387 unique genomic loci, was annotated onto the ENSEMBL genome browser. CGIs distribution strongly correlated with genome-wide gene density as previously shown (Lander et al., 2001). Furthermore gene rich T bands, proximal to the telomeres were also found to be overrepresented in the library (Craig and Bickmore, 1994). Interestingly, regions adjacent to the centromeres were shown to be relatively depleted for CGIs. This may represent unsequenced regions of the human genome, however this is unlikely as the CGI deficiency represents a larger proportion of the genome than that lacking tenable sequence (Bovee et al., 2008). It is feasible that this phenomenon results from large regions of pericentric heterochromatin, which are depleted for transcriptionally permissive sequences such as these.

CGIs were found to associate with 76.3% of protein coding genes, although surprisingly, mapping data suggested that approximately half of all CGIs are distal from gene promoters. It is possible that these 'orphan' CGIs represent as yet unidentified promoters (Gardiner-Garden and Frommer, 1987; Macleod et al., 1998). Indeed these may represent transcriptional initiation sites for ncRNAs such as *TSIX* and *AIR* which both originate from CGIs (Panning and Jaenisch, 1996; Sleutels et al., 2002; Wutz et al., 1997). Furthermore, non-promoter CGIs are enriched in chromosomal domains associated with homeobox genes. These complex gene loci contain multiple internal initiation sites and recent evidence has also highlighted the existence of functionally important antisense transcription within these regions (Kleinjan et al., 2004; Rinn et al., 2007). This will be discussed further in Chapter 4.

3.4.2 H3K4me3 and the origin of CGIs

One possible mechanism by which CGIs remain unmethylated, and by association, CpG rich, is steric hindrance of the *de novo* methyltransferases during global methylation in the early embryo (Reik, 2007). Since CGIs are known to colocalise with gene promoters, it is possible that steric hindrance could result from the association of the preinitiation complex at actively expressed genes in the developing embryo (Bird, 2002). This is consistent with the observation that genes with a tissue restricted expression profile which are expressed in the embryo possess a promoter CGI whilst others genes for which no transcript can be detected,

are associated with CpG deficient promoters (Daniels et al., 1997; Macleod et al., 1998). Indeed this is consistent with the induced *de novo* methylation and inactivation of the murine *APRT* gene resulting from mutation of Sp1 transcription factor binding sites (Brandeis et al., 1994; Macleod et al., 1994). This data presents the possibility that early embryonic transcription directly propagates and maintains CGIs in a nonmethylated form.

To test this hypothesis, we compared the CGI set with the pattern of the H3K4me3 in human ES cells (Guenther et al., 2007). This histone modification is known to associate with transcriptionally active promoters (Barski et al., 2007; Guenther et al., 2007; Weber et al., 2007). In accordance with this hypothesis, we discovered that 78.2% (13595 of 17387) of all CGIs colocalised to sites containing this histone modification. Furthermore, 63.5% (5574 of 8776) of the islands not overlapping an annotated TSS also mapped to this modification. This finding is consistent with CGIs arising from sites of transcriptional initiation in the developing embryo. It has been proposed that this histone modification is refractory to binding of *de novo* methyltransferases providing a possible mechanism by which the nonmethylated state of CGIs is maintained, even in the absence of somatic transcription (Ooi et al., 2007).

3.4.3 Comparing the CGI library with prediction algorithms

The CGI library was generated based on the empirical criterion of clusters of nonmethylated CpGs. As such it is unparalleled by sequence based prediction algorithms which do not account for DNA methylation status (Hackenberg et al., 2006; Ponger and Mouchiroud, 2002; Takai and Jones, 2002). The comparison of two commonly used predication methods illustrated this fact. Relaxed sequence parameters identify the majority of CGIs in the library, whilst greatly overestimating the actual number by more than ten fold. Conversely, restricting the parameters significantly reduced the overlap between the two sets but provides a much more accurate representation of the total complement of CGIs. However, 23% of CAP CGIs are not predicted in the stringent set, which is explained by the reduction of CpG and G+C composition of these sequences. Despite this, these islands are still coherent with respect to promoter association. This suggests that the prerequisite for lack of CpG methylation, allows CAP to identify islands which would otherwise fail to meet the sequence criteria for this particular sequence algorithm.

3.4.4 Summary

The CAP technique represents an effective preparative and analytical tool for the purification of CGI sequences from bulk genomic DNA. The utility of this approach was illustrated by the generation and characterization of a complete human somatic CGI set. This comprehensive library of islands bears sequence characteristics coherent with classical CGIs. Whilst providing useful information pertaining to genomic distribution and gene association of CGIs, the library also represents a useful analytical tool for further investigation of an interesting regulatory fraction of the human genome.

Chapter 4: CGI Methylation Analysis

4.1 Introduction

4.1.1 DNA methylation analysis

Initial studies investigating the role of DNA methylation were limited to determining the relative abundance of the methylated base within a given cell population. The chromatographic procedures applied had no spatial resolution with respect to the methyl base, but provided insight into the abundance and variability of DNA methylation *in Toto* (Ehrlich et al., 1982). However, the role of DNA methylation in a wide variety of biological processes has consequently led to the development of techniques to determine its distribution within a genomic context. These can be broadly separated into three functional categories based on the method of detection. These include methylation-sensitive restriction digestion, affinity purification and sodium bisulfite conversion. Each of these can be applied to DNA methylation profiling at specific candidate loci or across large genomic regions depending on the method used to resolve the information.

In mammalian systems, DNA methylation is restricted to the cytosine base in the context of CpG dinucleotides. Restriction endonucleases which are sensitive to CpG sites containing methylated (Hpa11, Hha1, BstU1, Not1 etc.) or unmethylated (McrBC) cytosine can be used to directly determine the methylation status of the recognition sequence. Early studies combined this selective digestion with Southern blotting to specifically determine the methylation status of cleavage sites within a defined DNA sequence (Bird and Southern, 1978). Whilst proving effective, this is a rather cumbersome technique and requires a relatively large quantity of starting DNA (approximately 10 μg per digestion). Alternatively, PCR amplification of DNA sequences which span methylation sensitive restriction cut sites can also be applied to infer methylation status and can be applied to nanogram quantities of DNA.

For global analysis, a combination of various restriction endonucleases coupled with cloning and sequencing, or microarray hybridisation have been described. Huang and coworkers developed a technology termed Differential Methylation hybridisation (DMH). This process employed combinatorial digestion using Mse1 and the methyl sensitive restriction enzymes

BstU1 and/or Hpa11. Methylated sequences resistant to restriction were selectively amplified by lmPCR (linker-mediated PCR) and probed against arrayed CGI amplicons, to identify cancer specific methylation events (Chen et al., 2003a; Huang et al., 1999; Yan et al., 2002; Yan et al., 2000). A similar technique was developed which employed the methyl sensitivity of Sma1 (CCCGGG) to identify differentially methylated CGI sequences. Determination of methylation state was achieved by DNA sequencing cloned inserts and subsequently by cohybridisation to CGI microarrays (Estecio et al., 2007; Heisler et al., 2005; Ueki et al., 2001). Restriction Landmark genome scanning, detects differential methylation between biological samples, by combining methyl-sensitive digestion with two dimensional gel electrophoresis. Restriction products, which are apparent in one sample but absent in another are determined visually, and identified by subsequent DNA sequencing (Plass et al., 1999; Yu et al., 2005). More recently, virtual RLGS profiles generated from standardized preparations and *in Silico* digestion have been used to expedite sequence identification (Matsuyama et al., 2003; Song et al., 2005). This procedure allows simultaneous interrogation of all sequences bearing the specific restriction site without the prerequisite sequence selection intrinsic to microarray analysis. Alternatively, specific restriction products can be size selected and cloned to generate a library of methylated DNA sequences. This was successfully applied to the identification of candidate islands that were methylated in Wilm's tumour samples (Strichman-Almashanu et al., 2002). Sequencing allows an unbiased detection of all methylated loci but is hindered by the presence of high copy number CpG rich repetitive elements such as the rDNA spacer (Brock and Bird, 1997; Strichman-Almashanu et al., 2002).

A further limitation of restriction techniques is the specific sensitivity of endonucleases to methylated DNA, which limits detection to nonmethylated sequences. Methylation is therefore indirectly inferred through an inability to cleave at the cognate recognition sequence, and is therefore more prone to systematic procedural artifacts. McrBC cleaves sequences containing the recognition motif (G/A^mC ($N_{40\text{-}3000}$) G/A^mC) when one or more cytosines are methylated. Combining the opposing sensitivities of Hpa11 and McrBC with PCR allows the direct determination of both methylated and nonmethylated sequences. Such increased resolution facilitated the detection of 3 monoallelically methylated CGIs in whole human blood DNA (Yamada et al., 2004). Schumacher and coworkers applied a similar technique, combining restriction with microarray hybridisation to characterise DNA methylation in a panel of brain samples (Schumacher et al., 2006).

Methyl-sensitive restriction is significantly limited by the fact that detection is confined to the available cleavage sites. This is highlighted for RLGS, where identification of differential methylation is limited to approximately 20% of human and mouse CGIs bearing Not1 recognition sites. Furthermore, only a subset of all CpG sites within a sequence context is assayable by this method. As such, methyl-sensitive restriction represents a quick but relatively low resolution technique for the characterisation of DNA methylation.

Alternatively, DNA sequences bearing methylated cytosine can be selectively enriched by various affinity purification methods. Currently, two reagents are available with specific affinity for methylated DNA. The first is the Methyl CpG Binding Domain, common to the MBD family of proteins (discussed in the introduction). Various recombinant MBD constructs from MeCP2 and MBD2 have been successfully applied to the specific recognition of methyl-cytosine in the context of CpG dinucleotides (Brock et al., 1999; Brock et al., 2001; Cross et al., 1994; Gebhard et al., 2006a; Gebhard et al., 2006b; Rauch et al., 2006). Alternatively detection can be achieved via a monoclonal antibody which recognizes the methyl-cytosine base in single stranded DNA (Keshet et al., 2006; Weber et al., 2005; Weber et al., 2007; Zhang et al., 2006; Zilberman et al., 2007).

MBD affinity can be used to selectively enrich for DNA sequences bearing clusters of methylated CpG sites from total genomic DNA. The MBD from MeCP2 and MBD2 have been used to enrich for methylated sequences which have then been screened for specific candidates of interest (Gebhard et al., 2006a; Shiraishi et al., 2002). Alternatively, methylated sequences can be cloned and sequenced ((Brock et al., 1999; Brock et al., 2001; Selker et al., 2003; Shiraishi et al., 1999)) or hybridised to a microarray of pre-selected sequences (Gebhard et al., 2006b; Rauch et al., 2006; Zhang et al., 2006). Specific details and applications of MBD affinity technologies will be discussed in the next section.

The utility of the methyl-cytosine antibody has been demonstrated in a number of global methylation studies. The method of immunoprecipitating DNA bearing the methyl mark is highly analogous to ChIP and is therefore well suited to microarray analysis. The initial study describing the technique, demonstrated the efficacy of the antibody for determining both global and CGI specific DNA methylation patterns in humans (Weber et al., 2005). It has subsequently been applied to the elucidation of DNA methylation profiles in normal (Weber et al., 2007) and cancerous (Keshet et al., 2006) human cells as well as the high resolution profiling of methylation in the *A. thaliana* genome (Zhang et al., 2006; Zilberman

et al., 2007). Methyl DNA Immunoprecipitation (MeDIP) has been applied to genome wide analysis in conjunction with microarray technology. However, PCR amplification of specific sequences from enriched DNA illustrates its utility as a candidate approach for DNA methylation analysis (Weber et al., 2005; Weber et al., 2007).

Presently little analysis has been carried out to directly compare the two reagents in terms of CpG density requirements, enrichment and selectivity. One study determined that the methylcytosine antibody could detect methylation at densities of two CpG sites per 100bp (Keshet et al., 2006). Whilst MBD technology is usually applied to the characterisation of sequences containing clusters of CpG sites, it remains unclear as to whether adjustment of the elution conditions would facilitate a similar level of sensitivity (Brock et al., 1999; Brock et al., 2001; Cross et al., 1994; Gebhard et al., 2006b). Indeed one comparison indicated that the two techniques gave highly concordant results in an analysis of the small genomed plant *Arabidopsis*. The antibody identified 20% of sites, not detected by MBD affinity, posited by the authors to indicate a difference in the CpG density requirements of the two reagents (Zhang et al., 2006). It is equally possible however, that this disparity arises from non-CpG cytosine methylation characteristic of many plants which would be undetectable using the MBD affinity method.

Affinity techniques are very powerful as they allow simultaneous DNA methylation analysis on a genome wide scale, limited only by the microarray platform utilised for detection. They do however fail to provide nucleotide resolution provided by techniques such as bisulfite genomic sequencing.

Sodium bisulfite converts all nonmethylated cytosine residues to uracil, but does not affect the methylcytosine moiety. Subsequent amplification of the converted DNA results in the conversion of uracil to thymine. As such, this method imprints a specific methylation state as a characteristic alteration in sequence composition. In its simplest form these changes can be identified by specific amplification of target regions followed by DNA sequencing (Frommer et al., 1992). This technique represents the 'gold-standard' in DNA methylation analysis as it provides single base resolution for every DNA strand in a mixed population. This was illustrated by high resolution mapping of DNA methylation spanning target regions on human chromosomes 6, 20 and 22. This provided detailed insight into tissue specific promoter methylation at a resolution unparalleled by previous studies (Eckhardt et al., 2004; Eckhardt et al., 2006). The cumbersome nature of the technique coupled with the cost of

deep sequencing presently limits its application to candidate methylation profiling. This can be partially alleviated by replacing conventional sequencing with Pyrosequencing, which provides a proportional representation of cytosine and thymine at each assayable CpG position (Yang et al., 2004). The reduced requirement for sequence reads is countered by a loss of single base resolution.

Related techniques exist which capitalize on the altered sequence composition resulting from bisulfite conversion. Combined bisulfite restriction analysis (COBRA) detects "methylation dependent sequence differences" by restriction digestion that discriminates between the altered sequences. This method allows a quick and quantitative measure of DNA methylation at a specific candidate locus (Xiong and Laird, 1997). Alternatively, coupling bisulfite conversion and various PCR based detection methods provides a measure of methylation status at pre-selected sequences. Methyl-sensitive single nucleotide primer extension (MS-SNuPE), employs specific primers which can be used to quantitatively assess the methylation status of individual CpG sites (Gonzalgo and Jones, 1997). This technique uses methylation state independent amplification and secondary detection methods to determine sequence composition and subsequently methylation state.

Alternatively, various techniques have been developed which determine methylation status by sequence specific oligonucleotide primers and probes. MethyLight employs a combination of methylation discriminatory primers and probes to quantitatively determine the methylation status of a specific DNA sequence (Eads et al., 2000). MethyLight is not applicable to global analysis but is suitable for determining the methylation status of specific sequences and repetitive DNA regions (Weisenberger et al., 2005). This is particularly informative in light of the fact that mammalian genomes are predominantly composed of DNA repeats, which are inaccessible by microarray and sequence based techniques. Methyl sensitive PCR (MSP) directly infers the methylation state by quantitation of amplicon produced from primers specific for the methylated and nonmethylated sequences (Herman et al., 1996). This technique is quick and simple and provides an effective means for confirming data from a global analysis such as those previously introduced.

As discussed bisulfite based methodologies can provide very high resolution information pertaining to DNA methylation within candidate regions, but is less suitable for genome wide profiling. To address this limitation various studies have developed microarray based bisulfite detections systems. The principle being to hybridise bisulfite treated DNA against

an oligonucleotide array containing specific probes for methylated and nonmethylated target sequences. A preliminary study showed that a focused array representing a portion of the human *ERα* region could effectively characterise the methylation status at nucleotide resolution (Gitan et al., 2002). However, to completely map CpG methylation at high resolution on a genome wide scale would require probes representing the methylated and unmethylated form of every CpG site in the genome. For human this would represent $6x10^6$ specific probes, with no feature replication. At present this represents a significant technical challenge but with the development of high density array platforms such as Illumina's bead arrays, this may represent a realistic goal in the future (Bibikova et al., 2006; Fan et al., 2006). Bisulfite sequencing is presently limited to focused genomic locations due to cost and the necessity for large numbers of region specific primers. However, the development of high throughput technologies such as solexa sequencing may allow the simultaneous characterisation of all unique CpG sites. Indeed the utility of this technology has already been indicated by the high resolution mapping of a range of post translational histone modifications (Barski et al., 2007). More recently, highly redundant sequencing of the small genomed plant, *Arabidopsis thaliana* confirmed that solexa sequencing could indeed be applied to the generation of global DNA methylation maps (Cokus et al., 2008).

Global sequencing technologies may provide the means by which entire genomic methylomes can be characterised. As such it is hard to imagine that such sequencing will not supercede array based technologies. However currently, cost and limited analysis tools present a restriction to nucleotide resolution mapping of mammalian genomes. Furthermore, it is widely accepted that polymerase based amplification is prone to sequence specific bias, and as such extensive characterisation will have to be carried out prior to the realisation of this potential.

4.1.2 MBD Affinity Purification (MAP)

MAP employs a recombinant methyl binding domain to selectively purify DNA fragments containing clusters of symmetrically methylated CpG sites. A recombinant MBD domain is covalently coupled to a chromatography matrix such that the binding affinity for methylated DNA sequences is preserved. The MBD matrix is subsequently applied to a suitable chromatography column to which fragmented DNA is bound under low stringency salt conditions. DNA fragments bearing unmethylated or low density CpG sites (1 per 100bp) can be eluted from the affinity matrix at low NaCl concentration. Subsequent application of an increasing NaCl gradient disrupts the electrostatic forces between the CpG rich

methylated DNA fragments and the coupled MBD domain (typically at a NaCl of > 700mM). This allows the efficient fractionation of DNA based on CpG density and methylation status. The affinity matrix is stable under these buffering conditions, allowing the elution of DNA whilst maintaining the functional integrity of the chromatography column for multiple purifications (Brock et al., 1999; Cross et al., 1994; Cross et al., 1999; Selker et al., 2003; Zhang et al., 2006).

MAP was originally developed for the selective purification of CGI fractions from bulk human genomic DNA (discussed in the previous chapter; (Cross et al., 1994)). This employed a two phased purification approach, with the initial stripping of CpG rich methylated DNA, and the subsequent enrichment of artificially methylated CGIs from bulk genomic DNA. The technique was also applied to the purification of CGIs from the genomes of mouse, pig and chicken in addition to targeted cloning of CGIs from human BACs, PACs and flow sorted chromosomes (Cross et al., 1999; Cross et al., 2000; Cross et al., 1997a; McQueen et al., 1997; McQueen et al., 1996).

More recently, MAP has been applied to the direct purification of methylated CpG rich fractions from bulk genomic DNA. The 'normal' methylation profile of DNA prepared from whole human blood was determined by sequencing a MAP enriched library. This identified a major fraction of methylated sequences derived from GC rich repetitive elements, and a unique population of sequences which localize to subtelomeric chromosomal termini (Brock et al., 1999). MAP, in conjunction with denaturing gel electrophoresis or subtractive hybridisation, was used to identify aberrantly methylated CGIs associated with lung Adenocarcinomas and breast cancer respectively (Brock et al., 2001; Shiraishi et al., 1999). The utility of MAP was particularly highlighted for the elucidation of methylation profiles in, as yet, unsequenced genomes. Isolation of methylated sequences from the filamentous fungus; *Neurospora crassa*, identified sequences which specifically localize to transposable elements. This data supported a role for DNA methylation in genome defense (Selker et al., 2003).

In addition to preparative applications, MAP can also provide a useful analytical tool for the direct inference of methylation state of pre-selected sequences. In the case of cancer samples, this application is particularly powerful as it allows methylation analysis to be carried out on nanogram quantities of DNA. This is a prerequisite when working with certain histological grade tumour biopsies where material is highly limiting. Candidate CGI methylation can be

determined directly by specific PCR amplification of sequences from MAP enriched DNA. This allows the quick, simultaneous assessment of target sequence methylation across a panel of biological samples (Gebhard et al., 2006a; Shiraishi et al., 2002). By extension, amplified, MAP-enriched DNA can be probed against a set of target sequences by microarray hybridisation. '^{Me}CpG Immunoprecipitation', a technique evolved from MAP, employs a bivalent MBD:IgG fusion protein and was used to screen aberrant CGI methylation events in Myeloid Leukemia (Gebhard et al., 2006b).

These studies indicate that MAP and its derivatives represent a versatile preparative and analytical tool for the elucidation of methylation events in a range of genomic contexts.

4.1.3 The Methyl-binding Domain

The MBD (pfam01429) is a short peptide motif which has specific affinity for symmetrically methylated CpG dinucleotides. It was initially identified as the DNA binding module in the mammalian corepressor protein; MeCP2 (Lewis et al., 1992; Nan et al., 1993). The amino acid sequence required for ^{me}CpG binding was determined through DNA binding assay using a panel of MeCP2 deletion mutants. This analysis identified a core sequence of approximately 85aa's which was sufficient for DNA binding *in vitro* (Nan et al., 1993). Cloning of the cDNA encoding the *MECP2* gene subsequently led to the identification of a family of proteins containing this highly conserved domain (Fig. 4.1-1; (Cross et al., 1997b; Hendrich and Bird, 1998)). In humans and mice there are five characterised members of this family, referred to as the MBD proteins. Of this family, four have intact DNA binding activity whilst MBD3 cannot bind DNA specifically (for details of the MBD proteins see the Introduction; Fig. 1.2-6; (Hendrich and Bird, 1998)).

Initial insight into the structural aspects of the MBD was provided by the NMR solution structure for the MBD of MBD1 in both the free and DNA complexed forms (Ohki et al., 2001; Ohki et al., 1999). On a gross scale, the MBD forms an amphipathic antiparallel α / β sandwich structure. This comprises of a four stranded β-sheet backed against an alpha helical turn which is stabilised by a core hydrophobic region (Ohki et al., 1999). The NMR structure of MeCP2 was found to be highly similar to that described for MBD1, adopting the same characteristic wedge shaped, α / β sandwich conformation (Wakefield et al., 1999). Both structures indicated that the COOH terminal loop in conjunction with the turn between β strands (2 and 3), protrude into the major groove to form one of the two DNA interacting interfaces (Ohki et al., 2001; Wakefield et al., 1999). A second DNA contact region occurs

via the amino terminal loop and a portion of the alpha helix. It is interesting to note, that despite the symmetry of the ^{m}CpG ligand, the MBD contacts the DNA through two distinct DNA interfaces (Ohki et al., 2001; Wakefield et al., 1999). Initial, characterisation of the MBD structures indicated that ^{m}CpG recognition was mediated by direct contacts via a conserved hydrophobic patch of amino acids. However, the X-ray crystal structure of the MBD of MeCP2 suggests that this recognition is indirectly coordinated by static water molecules. These are stably associated with the methylated cytosine base located in the major groove (Ho et al., 2008). This is supported by the fact that a conserved tyrosine residue, involved in water association, is mutated to phenylalanine in the mammalian MBD3 protein (Fig. 4.1-1; asterisked). Mutagenesis recapitulating the tyrosine at this position restores ^{me}CpG specific DNA binding (Fraga et al., 2003; Hendrich and Bird, 1998; Ho et al., 2008).

Interestingly, different *in vivo* binding specificities for the MBD family members were observed. This led to the identification of a preference for a run of A and Ts adjacent to the central methylated CpG site. It has been suggested that this introduces a bend in B form DNA, resulting in a significant narrowing of the minor groove which facilitates MBD binding (Ho et al., 2008; Klose et al., 2005). Footprinting analysis has confirmed that this results in the asymmetric binding of MeCP2 to its cognate CpG site. This specificity was encoded by the minimal MBD domain, but surprisingly was not observed for the highly similar MBD of MBD2 (Fig. 4.1-1;(Klose et al., 2005; Nan et al., 1993)).

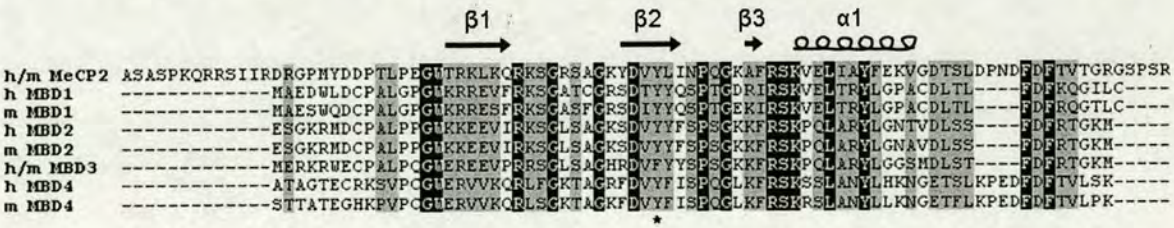


Figure 4.1-1. Conservation of the MBD in mouse and human
Amino acid sequence alignment reveals high levels of conservation between the MBDs of MBD family members. Regions of conservation are shaded, and the tyrosine to phenylalanine mutation in the MBD of mammalian MBD3 is asterisked. Species are denoted to the left of each sequence indicating human (h) and mouse (m). Secondary structure, as determined by the x-ray structure of the MeCP2 MBD, is indicated above the alignment with alpha helixes (black helix) and beta sheets (black arrows) indicated (figure adapted from (Ho et al., 2008)).

4.1.4 Differential CGI Methylation

Despite the fact that the large majority of CGIs are hypomethylated in all mammalian tissues, there are some which acquire methylation during cellular differentiation (Eckhardt et

al., 2006; Shen et al., 2007; Weber et al., 2007). A small but significant proportion of these islands is differentially methylated between tissues (Eckhardt et al., 2006; Futscher et al., 2002; Imamura et al., 2001; Kitamura et al., 2007; Nguyen et al., 2001; Oakes et al., 2007; Schilling and Rehli, 2007; Strichman-Almashanu et al., 2002). Several independent RLGS analyses identified tissue specific methylation in murine cells (Oakes et al., 2007; Shiota et al., 2002; Song et al., 2005). Approximately 5% of assayed sequences were found to be differentially methylated between tissues, although the majority of these occurred between testis and other somatic cells (Oakes et al., 2007; Song et al., 2005). An equivalent phenomenon was observed in an analysis of human brain, testis and Monocytes. Promoter arrays, probed with methylated DNA, enriched by methyl-cytosine immunoprecipitation, determined that the majority of differences resulted from testis specific hypomethylation (Schilling and Rehli, 2007). These results are consistent with the observation that CpG rich sequences which are methylated in somatic cells are frequently hypomethylated in mammalian sperm DNA (Brock et al., 1999). An independent microarray study identified CGIs which were methylated in human blood DNA. Candidate bisulfite analysis of seven of these islands confirmed that they were completely devoid of methylation in sperm DNA (Shen et al., 2007). However, despite the paucity of differentially methylated islands between somatic tissues, several such islands have been identified. The promoter region of *MASPIN* associates with a weak CGI sequence which was found to be differentially methylated in a panel of ten somatic tissues and cell types (Futscher et al., 2002). In rat tissues, a short (~200bp) differentially methylated region was identified in the 5' end of the promoter CGI of *SPHK1* (Imamura et al., 2001). A portion of the CGI associated with the promoter of human *SLC6A8* was identified as heavily methylated in four out of eight human somatic tissues (Grunau et al., 2000). More recently, bisulfite analysis of human chromosomes 6, 20 and 22, identified eleven CGIs which were differentially methylated between 8 somatic tissues (Eckhardt et al., 2006).

The increased representation of differentially methylated sequence between the germ line and somatic cells may have been selected for their ability to stably silence pluripotency genes (Zilberman, 2007). Somatic methylation of gene promoters transcribed specifically in cells of the germ line and embryo, could serve to irrevocably silence these genes in all somatic tissues. This is illustrated by the promoter methylation of the *MAGE* genes associated with transcriptional silencing in all somatic tissues (De Smet et al., 1999). Alternatively, such stable silencing may be too inflexible a mechanism to have a general role in the regulation of gene expression, between different somatic cell types (Zilberman, 2007).

4.1.5 Aim

Whilst extensive investigation pertaining to the role and distribution of DNA methylation in cancer has been carried out, relatively less attention has been paid to the equivalent features in 'normal' human tissues. Recent studies have provided tantalising insights into DNA methylation in human cells, although genome wide characterisation of tissue variability has yet to be carried out. To this end this chapter will investigate the CGI methylomes in a panel of primary human tissues utilizing a combination of MBD affinity purification and the novel CGI set introduced in the previous chapter. Whilst, it will be informative to characterise tissue specific genome wide methylation, this currently represents a technical challenge that has only been addressed for the small genomed plant *Arabidopsis thaliana* (Cokus et al., 2008; Zhang et al., 2006; Zilberman et al., 2007). We limit the analysis to CGIs as they represent a tractable fraction of the genome with obvious biological significance in both normal and disease processes.

4.2 Results: MBD affinity purification 'MAP' Array

4.2.1 Preparation of the MBD column

The MBD construct used in previous studies was generated by cloning the coding sequence of the rat MeCP2 MBD domain into the pet6h his tagging vector. The amino terminally tagged pet6hMBD construct proved to be an effective reagent for the affinity purification of methylated DNA (Cross et al., 1994). However, on expression, the majority of the protein was found to be insoluble and formed inclusion bodies (Bird lab unpublished observations). Despite attempts to increase solubility and expression[xxii], protein yields remained low. In order to make MAP column production more efficient, an alternative human MBD construct was generated.

The MBD coding sequence from human MeCP2 (amino acids 77-167) was amplified with PCR primers which introduced a 5' NdeI site (containing an in frame start codon) and a sequence corresponding to 6x histidine tag, two stop codons and an EcoR1 site at the 3' end.

[xxii] To increase expression and solubility, induced bacteria were grown for 5 hours at a reduced temperature of 30°C. In addition, IPTG and antibiotic concentrations were varied. Recombinant MBD was also purified under denaturing conditions, but the protocol is lengthy and resulted in low yields.

The amplicon was restricted with EcoR1 and Nde1 and cloned into the MCS of the pet30b expression vector (Fig. 4.2-1a). With the exception of the orientation of the histidine tag, the construct was essentially identical to the original MBD, containing the core amino acid sequence of the domain, encoding three conserved beta sheets and a single alpha helix (Fig. 4.2-1b; (Ho et al., 2008)).

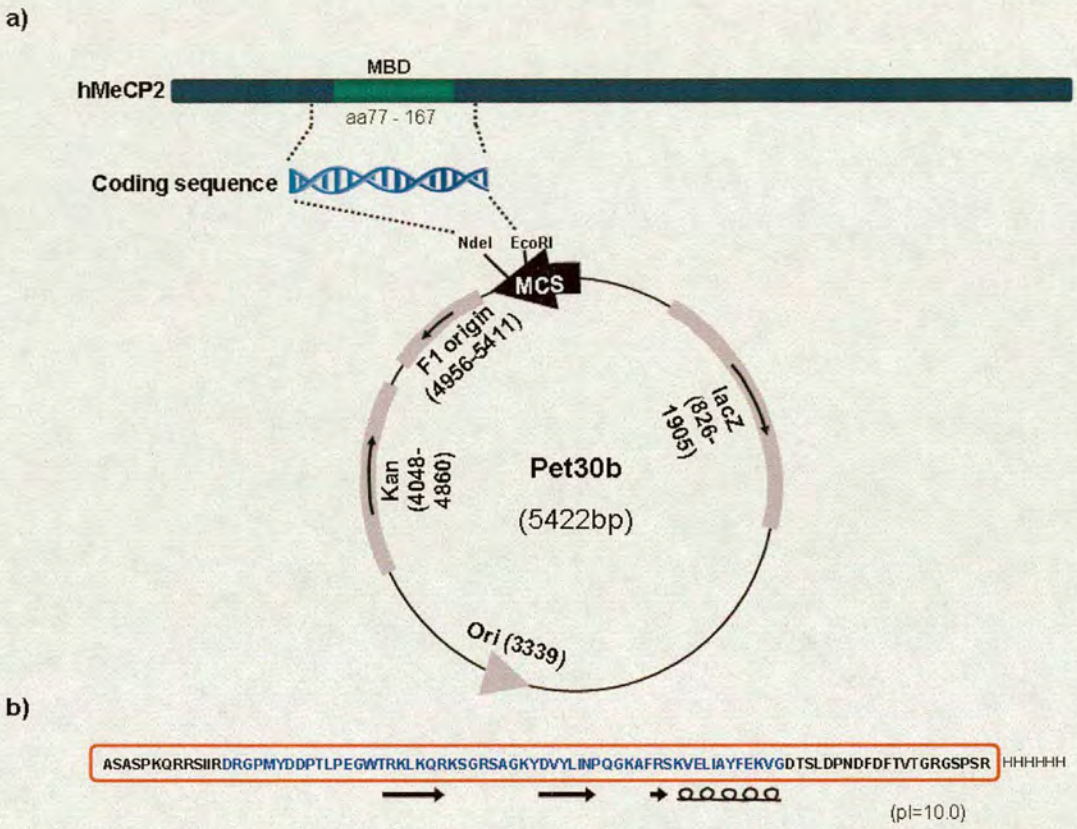


Figure 4.2-1. Cloning schema for the C terminally tagged MBD construct.
(a) Schematic representation of hMeCP2 indicating the position of the MBD domain (green). The location of the protein fragment (aas 77-167) is indicated. DNA encoding the MBD sequence was cloned into the Nde1 and EcorR1 sites of the pet30b protein tagging vector. The pet30b schematic is indicated as for Fig. 3.2-2a. **(b)** The MBD peptide sequence, indicating the location of the cloned amino acid sequence (red box) and the Histidine tag (6xH). Alpha helixes (black helix), beta sheets (black arrows) and minimal MBD sequence (blue lettering) are depicted.

Recombinant MBD was expressed by transforming chemically competent *E.coli* cells with the pet30bMBD construct. The MBD was bacterially expressed as previously described for the CXXC domain. SDS PAGE analysis indicated that IPTG treatment induced the expression of a protein corresponding to the size of the MBD construct (Fig. 4.2-2a). Furthermore, samples of the cell lysate confirmed that a large proportion of recombinant MBD was soluble (Fig. 4.2-2a, insoluble pellet (P) and soluble whole cell extract (WCE)).

Single step Nickel affinity chromatography efficiently removed contaminating proteins from the bacterial cell lysate without need for further purification (Fig. 4.2-2a).

To ensure that recombinant MBD recapitulated the methyl-specific affinity observed for the equivalent rat MBD construct, DNA binding was assayed by bandshift. This was carried out exactly as described for the CXXC construct, using both methylated and nonmethylated DNA probes to test for DNA binding. Consistent with previous findings, the MBD could efficiently bind to a probe containing methylated CpG sites but not to the same nonmethylated sequence (Fig. 4.2-2b). In addition, binding indicated that a single DNA probe could complex with multiple MBD domains. This suggested that an MBD affinity matrix would have a higher affinity for DNA containing clusters of CpGs as has been determined previously for both MBD and CXXC affinity techniques (Brock et al., 1999; Cross et al., 1994); Chapter 3).

Purified MBD was dialysed to remove the imidazole and then coupled to 1ml of Nickel charged sepharose. Saturated beads contained approximately 50mg of recombinant MBD. The affinity matrix was packed onto a 1ml chromatography column (Tricorn 5/50; GE Healthcare) and washed as described for the CXXC. Prior to calibration, the column was equilibrated with a 100ml, 0.1-1M, NaCl gradient to remove any contaminating bacterial DNA.

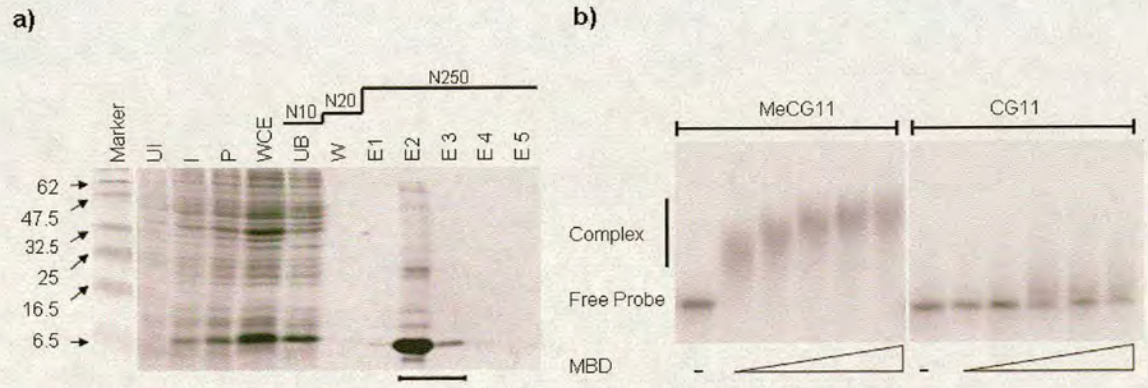


Figure 4.2-2. Purified MBD binds specifically to nonmethylated DNA.
(a) MBD induction and nickel affinity purification. Protein content of uninduced bacterial culture (UI), IPTG induced culture (I), insoluble lysate (P), WCE, unbound (UB), wash (W), elutions (E1-5) and size marker (Prestained Broad Range Marker, sizes indicated; NEB). Retained fractions are bracketed. **(b)** Bandshift assay showing the MBD complexed with a DNA probe containing 27 methylated CpG sites but not with a probe containing no methylated CpGs. Nonmethylated probe DNA (CG11) or methylated probe (MeCG11) was incubated with 0, 25, 50, 125, 250 or 500ng of recombinant MBD protein.

4.2.2 MAP Calibration

MBD chromatography has been shown to specifically purify DNA sequences containing clusters of methylated CpG sites (Brock et al., 1999; Cross et al., 1994). However, variations in the column matrix, MBD construct and the utility of the technique necessitated thorough calibration prior to its application. For this, a preliminary plasmid fragment calibration was carried out as previously described for the CXXC.

500ng of methylated and nonmethylated pABS fragments were prepared and end labeled as described in the previous chapter (depicted in figure 3.2-5a). Nonmethylated plasmid fragments were bound to the MBD column in low salt (0.1M NaCl) and eluted across a 0.1-1M NaCl gradient. This process was repeated for the methylated DNA fragments. The DNA content of each fraction was measured using a liquid scintillation counter (Perkin Elmer). Fractions were then resolved by agarose gel electrophoresis to determine the specific MAP elution profile of methylated and nonmethylated DNA fragments (Fig. 4.2-3a).

As expected the MBD column had reduced affinity for nonmethylated DNA fragments relative to their methylated counterparts. All three nonmethylated fragments eluted at approximately 750mM NaCl concentration irrespective of CpG density. Methylated DNA fragments however, required higher salt concentration for elution (>850mM NaCl). As expected, the smaller CpG deficient fragment eluted at a lower NaCl concentration (850mM) than the two larger, more CpG rich sequences (retained until 900mM NaCl; Fig. 4.2-3a).

This plasmid fragment calibration indicated that MAP could enrich for artificially methylated CpG rich DNA from a low complexity pool of DNA fragments. However, in contrast to the resolution obtained with CAP, the nonmethylated and methylated sequences were distinguished by as little as 200mM NaCl concentration. In order to ensure that this separation was sufficient to resolve endogenous methylated CGIs from bulk genomic DNA, further calibration was required.

Human male and female blood genomic DNA (n=3) was restricted with the endonuclease MseI (TTAA) and fractionated over the MBD column. The salt gradient was optimized to incorporate a wash at 750mM NaCl concentration to increase the resolution between the nonmethylated and methylated DNA fragments (Fig. 4.2-3b). This NaCl concentration was chosen based on the preliminary MAP calibration which indicated this to be sufficient to

remove all nonmethylated DNA. Relative DNA content[xxiii] of MAP fractions indicated that bulk genomic DNA was eluted from the affinity matrix at a NaCl concentration of 650-700mM (Fig. 4.2-3b). PCR was applied to investigate the elution profile in more detail by specific amplification of a range of CGI sequences with known methylation status. Nonmethylated CGIs (*P48* and *XIST* on the Xi in females) co-fractionated with bulk genomic DNA, whilst methylated CGIs (*NYESO* and *XIST* in males and Xa in females) eluted at a higher NaCl concentrations (800-1000mM). Furthermore, it is interesting to note that the *NYESO* CGI (2168bp containing 96 CpG sites) bound to the MBD affinity matrix more tightly than the short *XIST* CGI fragment (214bp containing 10 CpG sites). This confirmed the requirement of multiple CpGs for efficient retention by the MBD affinity matrix (Fig. 4.2-3b).

This calibration indicated that the affinity matrix had selective affinity for methylated DNA fragments. Furthermore, although elevated CpG density was not required for efficient purification *per se*, MseI digestion of CpG deficient bulk genomic DNA would result in DNA sequences with insufficient CpG sites to bind the MBD matrix. The resolution of MAP is poorer than that of CAP, however introduction of a NaCl wash step at 750mM effectively separated methylated CpG rich sequences from bulk genomic DNA. The genomic calibration was repeated for each new MBD column constructed to account for variations in recombinant MBD and matrix packing density. The elution profiles for each column were indistinguishable from one another, although the exact NaCl concentration at which DNA fragments eluted from the column varied by as much as 40mM. To account for this, the wash step was adjusted to an appropriate salt concentration as defined by bulk genomic elution and the affinity of a panel of specific DNA sequences. This adjustment was necessary to allow consistent sample preparation for microarray hybridization.

4.2.3 MBD column stability

MAP fractionations were carried out over the course of more than twelve months. During this time, four MBD columns were constructed, each of which was individually calibrated as outlined. After approximately fifty fractionations, repeat calibration indicated reduced DNA binding to the MBD affinity matrix (data not shown). This temporal deterioration could have resulted from the reduction of the nickel beads, protein degradation or stripping of nickel ions. However, it is unlikely to have been reduction of the Ni^{2+} ions as this would have

[xxiii] Relative DNA content was determined by measuring UV light absorbance at 260nM using the chromatography apparatus.

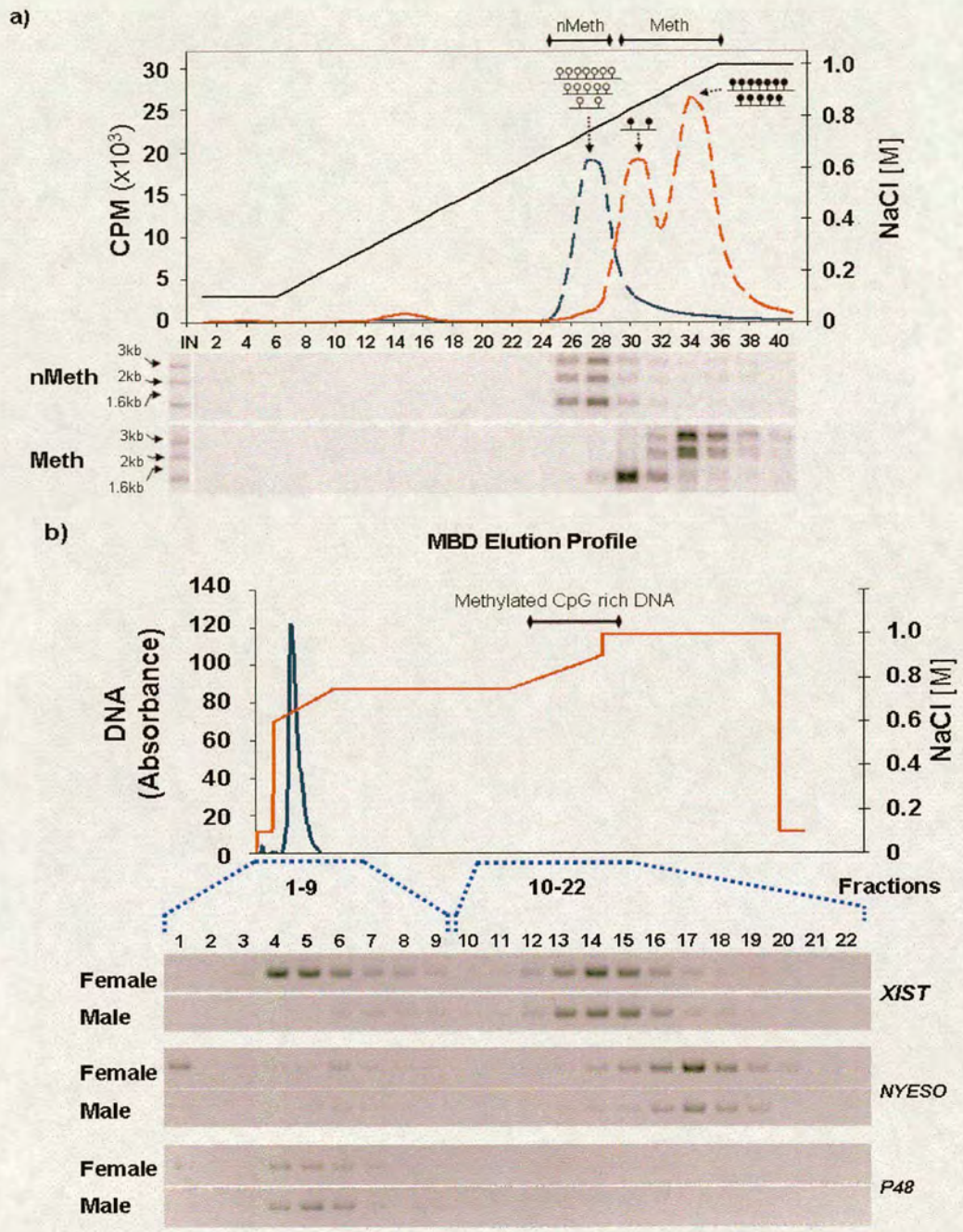


Figure 4.2-3. MAP preferentially binds to methylated CpG rich DNA.
(a) The upper panel indicates the elution profile of nonmethylated (dashed blue line) and methylated (dashed red line) DNA fragments across a NaCl gradient (solid black line). All nonmethylated fragments (nMeth) elute at approximately 0.75M NaCl whereas methylated fragments (Meth) don't elute until 0.85M with the most CpG dense fragments being retained by the CXXC matrix until 0.95M. Figure layout is depicted as for Fig. 3.2-5b (**b**) MAP fractionation of digested male and female genomic DNA. The upper panel is a representative elution profile of bulk genomic DNA (blue line). Genomic DNA (50µg) was applied to the MBD affinity matrix (see Methods) in low salt (0.1M NaCl) and eluted over an increasing NaCl gradient (red line). Twenty two fractions (dashed blue lines) were interrogated by PCR. The bracket above indicates fractions containing methylated CGIs. The lower panel indicates the elution profile of specific CGI sequences of known methylation status. Nonmethylated CGIs (*P48* and *XIST* in females) coelute with bulk genomic DNA, whereas the methylated CGIs (*NYESO* and *XIST*) elute at high NaCl concentration.

resulted in the formation of a brown colour which was not observed. Imidazole elution of the MBD from the Nickel sepharose matrix indicated no appreciable protein degradation after a period of six months (Fig. 4.2-4a). Nickel stripping is the most likely cause of the observed deterioration as this is known to affect metal affinity cation exchange, and is usually countered by recharging the matrix with nickel. As this was not an option, alternative matrix chemistry was applied in an attempt to stabilize MBD adsorption.

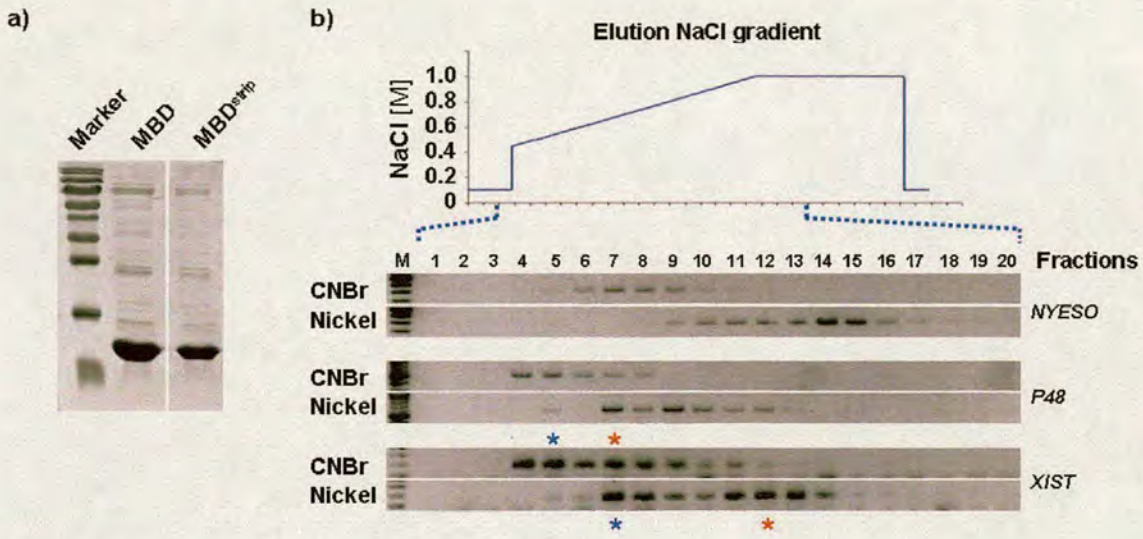


Figure 4.2-4. Optimisation of the MBD affinity matrix.
(a) SDS PAGE analysis indicated that MBD protein stripped from nickel sepharose after six months of chromatography (MBD^{strip}) showed little sign of degradation relative to a freshly purified batch of the MBD (MBD). Prestained Broad Range Markers 6-175kDa are indicated (Broad Range Marker; NEB). (**b**) CNBr coupling reduces the affinity and resolution of the MAP procedure. Fractions from CNBr and Nickel MBD column purification of female human blood DNA were assayed for specific DNA sequences by semi-quantitative PCR. The CNBr column showed ^{me}CpG specific biding as confirmed by the increased affinity for the CGI of *NYESO* (methylated) relative to that of *P48* (nonmethylated). Despite this, the resolution between methylated and nonmethylated DNA was greatly reduced relative to the nickel equivalent. The nonmethylated and methylated peaks of *XIST* (asterisked; blue and red respectively) were separated with less resolution by the CNBr column. Fractions interrogated by PCR are indicated (dashed blue line).

Activated CNBr forms covalent attachments with primary amino groups, facilitating the stable coupling of proteins to a sepharose matrix. To investigate the effect of this adsorption technique, 40mg of purified MBD was coupled to either 1ml of CNBr sepharose or 1ml of Nickel sepharose. Adsorption efficiency was equivalent for both matrices as determined by Bradford analysis of input, wash and unbound fractions. To test MAP functionality, 7.5 µgs of Mse1 digested female genomic DNA was applied to each column and chromatographed as previously described. Fractions were subsequently interrogated by PCR amplification of control DNA sequences. The Ni-MBD column behaved as previously characterized, however the CNBr-MBD column showed reduced affinity for all DNA sequences tested.

Methylated CpG dense fragments were bound preferentially, but the resolution of separation was deemed insufficient for effective purification of methylated DNA (Fig. 4.2-4b; compare blue and red asterisks).

It is possible that the reduced binding affinity and specificity of the CNBr matrix was due to excessive multi-point coupling of the MBD. This would have resulted in the occlusion of the active DNA binding site. Therefore, optimisation of binding conditions could yield a functionally equivalent column that is more stable than the chemistry currently employed. Future investigation will be required to determine if this is the case, however for the purpose of this study, the generic Nickel matrix was applied successfully. Deterioration associated with this type of column was countered by frequent recalibration, and a maximum functional lifespan of 40 MAP runs per column.

4.2.4 CGI microarray platform

Whilst it would be highly informative to investigate genome-wide methylation patterns in human tissues this is currently beyond our technical abilities. Techniques such as MAP and MeDIP are not sensitive to DNA methylation at the CpG density (1 site per 100bp) characteristic of bulk genomic DNA (Brock et al., 1999; Cross et al., 1994; Weber et al., 2005). Furthermore, bisulfite sequencing techniques are not, as yet, applicable to large mammalian genomes (Eckhardt et al., 2006). As such this study focuses on the DNA methylation status of CGIs which represent an interesting and tractable fraction of the human genome, the relevance of which has previously been discussed. Having established the MAP chromatography procedure for the fractionation of methylated DNA sequences, global analysis required a suitable microarray for CGI identification.

The microarray was generated from DNA sequences present in the CAP CGI library characterised in chapter 3. Amino-link amplicons were prepared by PCR amplification of CGI clone inserts with a 5' aminolink oligonucleotide forward primer. Amplicons were spotted and covalently coupled to the surface of amine binding glass slides (for full details see http://www.sanger.ac.uk/Projects/Microarrays/arraylab/methods/shtml). In total 28,800 features were spotted onto the arrays, representing 27,666 clone inserts (including duplicates) and 1134 negative control spots. 59% of the 17,387 unique CGIs were represented by duplicates which represent a combination of equivalent sequence clones and contiguous Mse1 fragment inserts from the same island.

During microarray data analysis, duplicates were not merged to give additional technical replication due to the fact that not all CGI were represented multiple times. Furthermore the two distinct classes of duplicate complicated their utility. However these features provided a useful empirical means of verification for all aspects of microarray data analysis. The microarray was fabricated entirely by Cordelia Langford and colleagues at the Wellcome Trust Sanger Institute Microarray facility.

This microarray represents approximately 60%[xxiv] of all CGIs in the human genome which is comprehensive when compared with equivalent resources (Heisler et al., 2005). However, in the future an exhaustive set of CGIs representing the entire human genomic complement would be beneficial. Furthermore, due to the fact that CAP only enriches for CGIs which are nonmethylated in blood DNA, those which are fully methylated will necessarily be absent from the library (Eckhardt et al., 2006; Weber et al., 2007). This is illustrated for the promoter CGI of the cancer testis antigen gene *NYESO*, which is fully methylated and therefore not enriched by CAP (Fig. 3.2-6a; Fig. 4.2-3b; (De Smet et al., 1999)). However, despite the absence of these sequences, the array allows the detection of CGIs showing fractional methylation or indeed differential methylation between somatic tissues. Completion of the CGI set will be discussed later.

4.2.5 Procedure for MAP array and data normalisation

A total of 50µgs of genomic DNA was pooled from three individuals; MseI digested and ligated to adaptors suitable for lmPCR. The prepared DNA was fractionated by MAP, and sequences retained with high affinity were re-chromatographed for a second time. MAP enriched and input DNAs were PCR amplified and fluorescently labeled by random incorporation of Cyanine 3 or 5 conjugated dCTP using the Klenow DNA polymerase fragment. Alternately labeled Input and MAP DNA was then cohybridised to the CGI microarray (Fig. 4.2-5).

DNA from each biological sample was prepared and MAP fractionated in duplicate. The MAP probes were then amplified and labeled in both Cy3 and 5 orientations to allow for further technical replication (dye swap experiments). The four technical replications were introduced to reduce experimental artifacts resulting from cyanine incorporation bias or variations in MAP preparations. The microarrays were scanned on an Axon 4200B

[xxiv] 60% coverage was estimated based on a total genomic complement of 25,500 CGI, the justification for which was discussed in chapter 3.

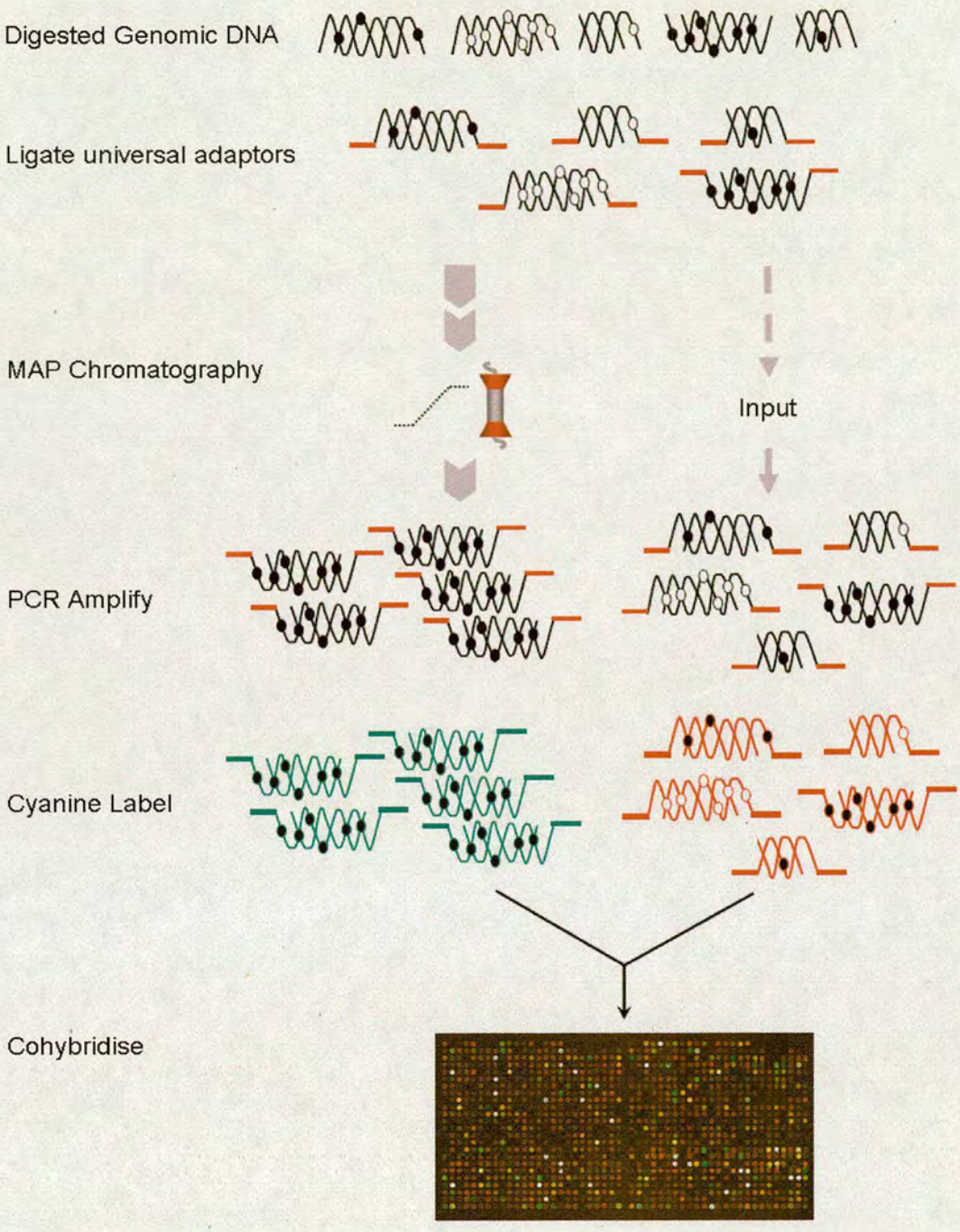


Figure 4.2-5. Schematic showing the isolation of methylated CpG dense DNA sequences form bulk genomic DNA. Nonmethylated (open circles) and methylated CpGs (filled circles) are indicated. Red bars denote ligated universal adaptors, and green and red lines represent cyanine 3 and cyanine 5 labeled DNA respectively.

autoloader and the raw signal intensities and primary quality controls[xxv] were carried out using the Genepix pro software package (Axon). Raw signal intensity values for each

[xxv] The primary data generated by the microarray scanner were TIFF images off raw Cy3 and Cy5 spot intensities. Each image was visually inspected to determine 'bad' spots and then auto flagged to identify weak or absent features. These flag values were then introduced into later stages of analysis to down weight or remove 'bad' features.

replicate microarray were calculated and then analysed using the limma (Linear Models for Microarray Data) package and R statistical environment (Smyth and Speed, 2003). Feature intensities were corrected for local background signal and those with very low values (≤1000) were coerced to a minimum value of 1000[xxvi]. Relative DNA quantities for all quality controlled CGI feature were calculated as a $\log_2$ ratios of Cy5:Cy3 (M value), required for further analysis. To remove systematic trends occurring from technical aspects of the microarray experiment, the M values were normalized using print tip loess (Smyth and Speed, 2003). This fitted the data to a local regression line centering the data on a log ratio of zero. A typical data set and the processing involved is illustrated for quadruplicate hybridisations of MAP prepared genomic DNA (Fig. 4.2-6). Mean M values across all replicate experiments were extrapolated using a linear model fit and then used to determine the methylation status of each CGI feature (see Materials and Methods for full details).

The library represented on the microarray contains a proportion of CGI amplicons which are indicative of CGIs but which only contain the periphery of the whole island. As such these inserts were informative for the identification and localisation of CGIs but contain too few CpG sites to be efficiently retained by the MBD affinity matrix under the elution conditions applied. As such all CGI inserts with a $CpG_{[o/e]}$ value of less than 0.5 were removed from further analysis. Furthermore, sex chromosomes were trimmed from the dataset (unless otherwise stated) to account for the fact that pooled individuals were not all of equivalent sex. Methylation analysis, subsequent to the calculation of M values, was carried out on this trimmed set representing 14,318 CGI inserts (Dataset 2).

4.2.6 CGI methylation in Male and Female blood DNA

A pilot experiment was carried out to characterise the CGI methylation profile in a human somatic tissue and to determine the efficacy of the MAP array methodology. Human genomic DNA was prepared from male and female whole blood and MAP fractionated as previously described. Two independent replicates for each sex were prepared, and these in turn were each hybridised as two separate dye-swap pairs to give four individual replicates. As expected, the mean M values for each sex were positively correlated (R = 0.865: Pearson correlation; Fig. 4.2-7a).

[xxvi] Microarray experiments generally produce large numbers of features with low signal intensity values which introduce artificial variation between hybridisations. As such, minimum capping such as this stabilises the data set and provides more reliable results.

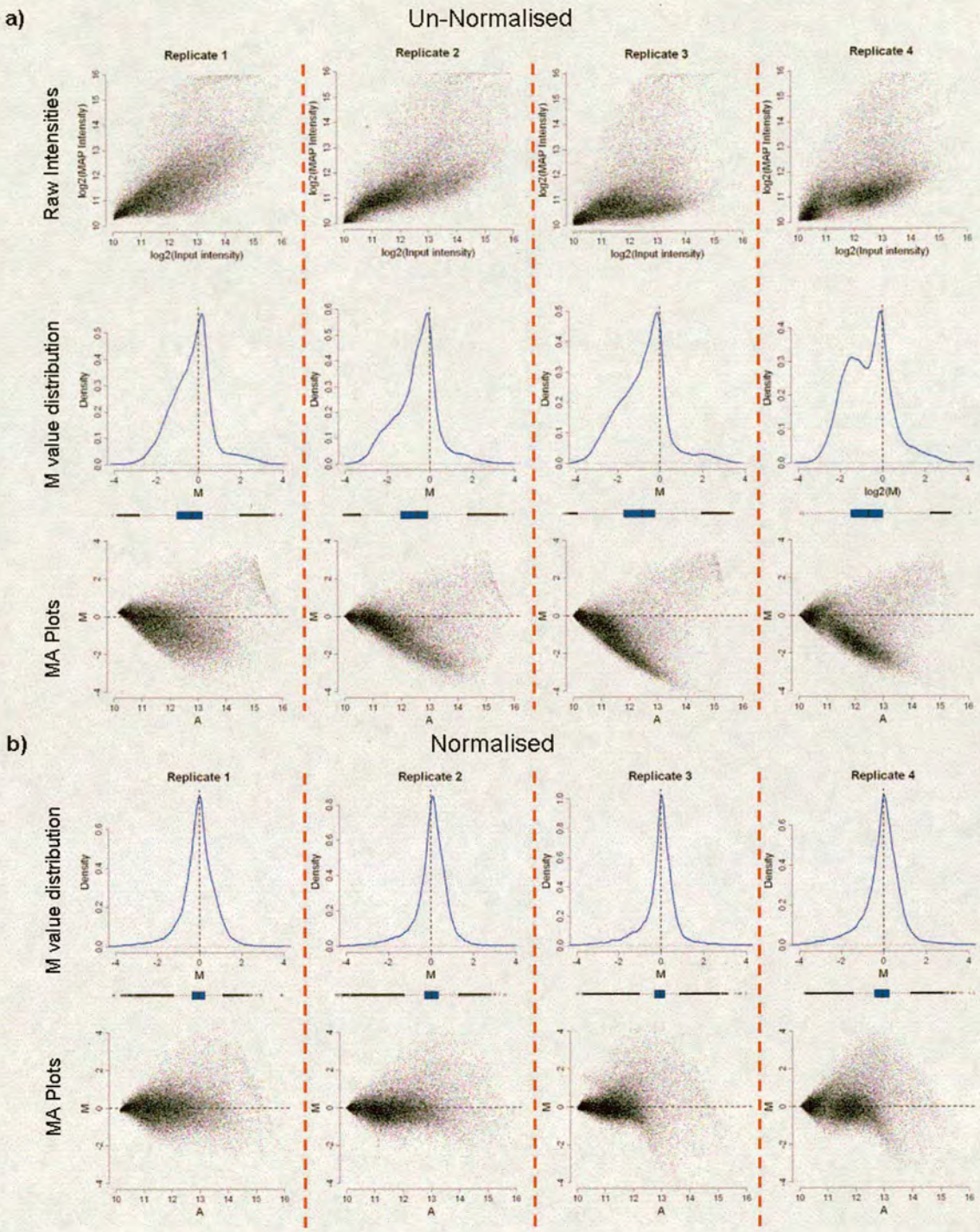


Figure 4.2-6. Replication and normalisation of MAP array data.
(a) Un-Normalised intensity data from a MAP array experiment. Plots of Raw signal intensities, M value distribution and M vs. A (measure of intensity) highlight the variability between quadruplicate experiments (Replicates 1-4). Furthermore, the plots indicate the fact that the majority of features present on the microarray have higher input intensity values relative to MAP purified DNA. This is consistent with the majority of CGIs being unmethylated and therefore not retained by the MBD affinity matrix. A $\log_2$ Cy3:Cy5 ratio of zero (dashed black line) is indicated. Boxplots are represented as per Fig. 3.1-1 **(b)** Print tip Loess normalisation centers the data and improves the consistency between replicate experiments. Comparison of M value distribution, as highlighted by the boxplots in **(a)** and **(b)** (as presented in Fig. 3.1-1) indicates the refined distribution of the data by the normalisation process.

One of the best characterised examples of CGI methylation is that associated with the inactive X chromosome (Heard et al., 1997). As such it was expected that the majority of sex specific differences in methylation would affect X-linked CGIs (Weber et al., 2005). To investigate this, the relative enrichments for all CGIs on the X chromosome were compared with those on chromosome 16. The autosome showed no sex specific difference in methylation status, whereas a significant proportion of all of the X-linked CGIs showed elevated M values in female relative to male (p value < 2.2×10^{-16}; Wilcoxen Rank Sum Test; Fig. 4.2-7b-d). This comparison was extended, confirming that whilst methylation profiles were indistinguishable between the two sexes for all twenty two autosomes, methylation levels of CGIs on the X chromosome were elevated in females relative to males (p value < 2.2×10^{-16}: Wilcoxen Rank Sum Test, Fig. 4.2-7d). It is interesting to note that specific Y-linked CGIs are also enriched in females relative to males despite this being significantly hindered by the lack of this particular chromosome (Fig. 4.2-7d). The X and Y chromosomes are derived from the same ancestral autosome, and despite the evolutionary decay of the latter, it still bears significant sequence homology to its ancestral partner (Graves, 2006). This suggests that this result is an artifact resulting from cross hybridisation of methylated X-linked CGIs with homologous regions on the Y chromosome.

The process of sex chromosome dosage compensation in mammals, extinguishes the expression of the majority of genes on one of the two X chromosomes. A notable exception is the *XIST* transcript, a functionally important ncRNA, which is specifically expressed from the inactive X chromosome. *XIST* activity is inversely correlated with the methylation status of its promoter CGI which has been implicated to have a role in its repression ((Panning and Jaenisch, 1996), Fig. 4.2-3b). However *XIST* is not the only gene found to escape X inactivation. Studies in human and mouse have identified a panel of X-linked genes which are biallelically expressed (Carrel et al., 1996; Carrel and Willard, 2005).

The negative correlation between *XIST* expression and promoter methylation coupled with the fact that promoter CGI methylation prevents transcription directed the investigation of methylation status of these biallelically expressed genes (Hansen and Gartler, 1990; Panning and Jaenisch, 1996; Stein et al., 1982). The methylation status of X-linked CGIs associated with the promoters[xxvii] of biallelically (n=14) and monoallelically expressed (n=103) genes was compared. As expected, MAP probes representing the inactivated class of CGIs had an

[xxvii] ± 1.5kb of an annotated transcription start site (TSS).

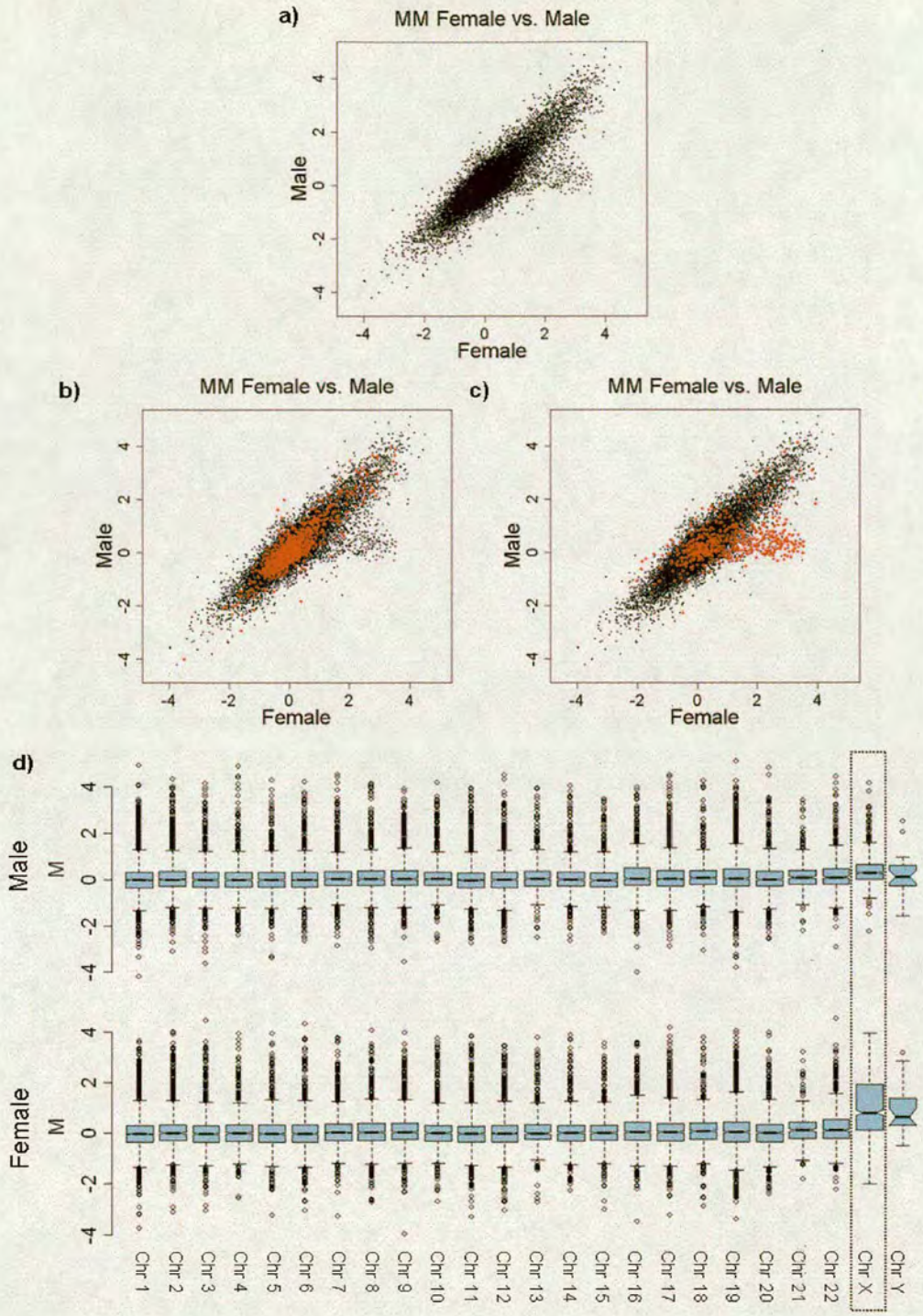


Figure 4.2-7. MAP array identifies female specific CGI methylation on the X chromosome

(**a-c**) Plots indicating the M values of arrayed CGIs MAP enriched from male and female whole blood genomic DNA. (**a**) The enrichment profile of all CGI probes isolated by MAP from female and male genomic DNA are positively correlated (R = 0.865; Pearson correlation). (**b**) CGIs present on chromosome 16 (red dots) show a methylation status indistinguishable between male and female MAP preparations. (**c**) MAP probes isolated from male and female genomic DNA indicate female specific methylation of X-linked CGIs (red dots). (**d**) Boxplots depicting the M values for CGIs representative of all male (top panel) and female (lower panel) chromosomes indicate X-linked, female specific, CGI methylation. Boxplots are depicted as per Fig. 3.1-1 with X chromosome highlighted (dashed box).

M value distribution indistinguishable from that of all X-linked CGIs (Fig. 4.2-7d and Fig. 4.2-8). Conversely, promoter CGIs which escape X inactivation were found to be distinct in that they were significantly less enriched (KS test: p.value = 1.2X10-5; Fig. 4.2-8). This suggested that CGI genes which escape X inactivation have nonmethylated CGIs which are permissive to transcription. The specific complement of biallelically expressed genes varies between females (Carrel and Willard, 2005), and may represent uncharacterised epigenetic polymorphisms with a role in individual and sexual dimorphic traits.

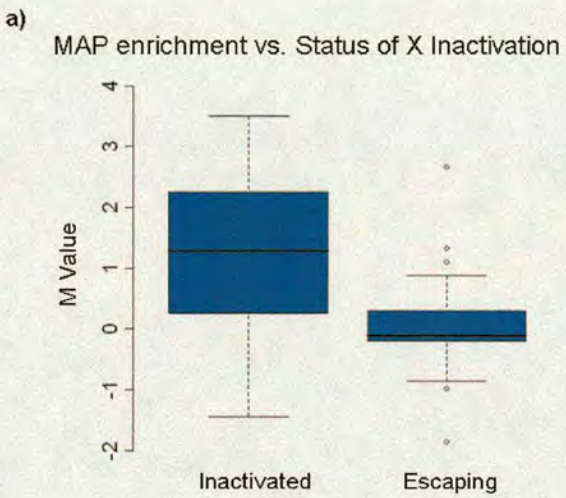


Figure 4.2-8. Relationship between promoter CGI methylation and gene expression on the inactive X chromosome.
MAP array distinguishes promoter CGIs associated with genes which undergo (Inactivated) or escape (Escaping) X inactivation. X-linked genes, previously shown to be repressed on the inactive X chromosome in females (n=103), are significantly more enriched by MAP relative to those expressed from both X chromosomes (n=14, KS test; $p = 1.2 \times 10^{-5}$).

To determine the minimum threshold value which describes CGI methylation, 12 CGI inserts spanning a range of M values were amplified from MAP and input DNA (from the female purification). Results from triplicate quantitative PCR analysis indicated that an M value of 1.5 could identify methylated CGIs (Fig. 4.2-9a). Bisulfite genomic sequencing was carried out on two randomly selected islands from male and female genomic DNA. The CGI I3878[xxviii] associated with the *BEST* gene had M values of 0.2 and 0.4 in female and male hybridisations respectively and, as predicted, was confirmed to be unmethylated by bisulfite genomic sequencing (Fig. 4.2-9b). CGI I19112 spans the *R4RL1* coding region and has an M value in excess of 2. When investigated, this island was found to be heavily methylated in both male and female blood DNA (Fig. 4.2-9c). In combination with additional bisulfite

[xxviii] CGI nomenclature used throughout this chapter correspond to the microarray feature identifiers.

genomic sequencing results these findings suggest that an M value of ≥1.5 was suitable for the identification of methylated CGIs (data discussed in the next section). The threshold applied to describe CGI methylation is relatively stringent, and as such likely excludes a proportion of *bona fide* methylated islands. However, for the purpose of this study constricting the identification of methylated islands served to reduce false positive.

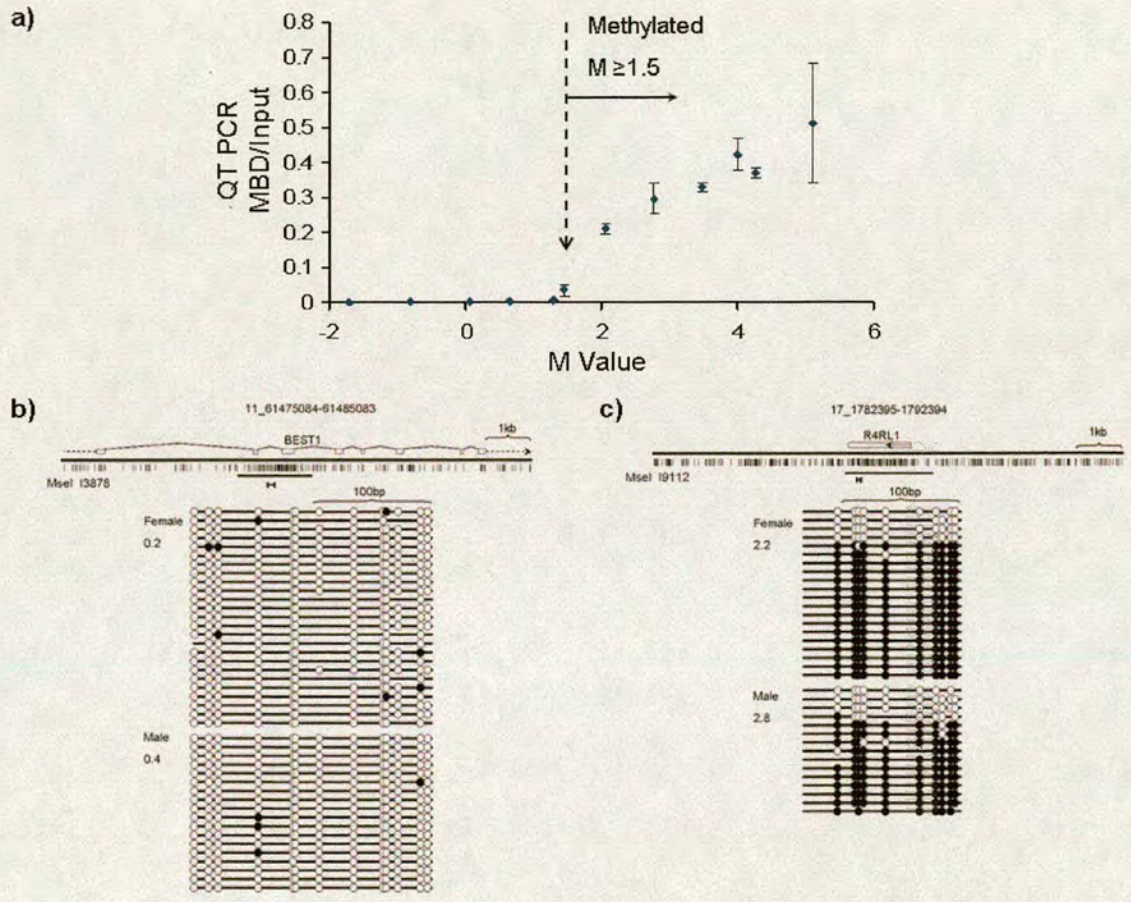


Figure 4.2-9. Validation of minimum M value required to determine methylation status.

(a) QT PCR analysis of MAP enrichment for CGI fragments with a range of M values indicates appreciable enrichment with M values equal to or greater than 1.5. Error bars represent standard deviation values (SDs) across three replicate experiments. **(b and c)** Confirmation of methylation status indicated by MAP array by bisulfite genomic sequencing. CGI inserts I3878 **(b)** and I9112 **(c)** were confirmed to be nonmethylated and methylated respectively, as predicted by the MAP array results. Open and filled circles represent non-methylated and methylated CpG sites respectively. The genomic locus including annotated transcripts and CpG sites (vertical strokes) are shown above each profile. Each column represents products of amplification by a single primer pair (location indicated by brackets below CpG map). Each line corresponds to a sequenced DNA strand. Black bars indicate the location of the MseI fragment cloned in the CGI library.

The application of these criteria (p value ≤ 0.01[xxix], M value of ≥ 1.5) to the male and female hybridisation data, confirmed extensive methylation of X-linked CGIs. In contrast, the percentage of CGIs methylated on the single male X chromosome was comparable to the levels found on the human autosomes, as illustrated for Chromosome 16 (Table 4.2-1). In addition, it is clear from this data that not all autosomal CGIs are nonmethylated, which is consistent with previous studies (Eckhardt et al., 2006; Shiota, 2004; Shiraishi et al., 2004; Song et al., 2005; Strichman-Almashanu et al., 2002; Weber et al., 2005; Weber et al., 2007; Yamada et al., 2004). These results confirmed the effective utility of MAP array to identify methylated CGIs from bulk genomic DNA.

Table 4.2-1. Methylated CGIs on Chr16 and ChrX in Human Whole Blood DNA

Chromosome	Sex	Total[a]	Methylated Islands	Methylated Islands (%)
16	Male	569	49	8.6
16	Female	569	62	10.9
X	Male	357	32	9.0
X	Female	357	151	42.3

[a]Number of CGIs based on the CpG density filtered set (discussed in the previous section)

4.3 Results: Global Human DNA methylation analysis

4.3.1 Tissue specific CGI methylation

Human cellular processes such as genomic imprinting and X-inactivation are known to associate with the methylation of CGIs. Further to these well characterised examples an increasing number of methylated islands have been identified in other genomic locations (Eckhardt et al., 2006; Futscher et al., 2002; Song et al., 2005; Weber et al., 2007; Yamada et al., 2004). In the previous section, MAP array was applied to the identification of methylated islands in whole blood genomic DNA. These experiments identified 151 (42.3%) of CGIs which were methylated on the X chromosome which was consistent with female X inactivation. However a further 910 (6.4%) were identified as methylated on the twenty two autosomes. It is unclear, whether these CGIs represent a set of ubiquitously methylated islands present in all human tissues or whether there is appreciable variability between cell types, consistent with a potential role in tissue specification and gene regulation. Despite this

xxix Significance was determined by linear modeling of the data, followed by the application of an empirical Bayes (eBayes) method. Benjamin Hochberg adjustment was applied for multiple testing correction to provide a more stringent measure of significant enrichment.

however, the global function and variability of CGI methylation are still poorly defined (Eckhardt et al., 2006; Futscher et al., 2002; Shi et al., 2002; Song et al., 2005; Weber et al., 2007; Yamada et al., 2004).

To address these uncertainties, MAP array probes were prepared from germ line (sperm) and three somatic (Brain, Skeletal Muscle and Spleen) DNAs and hybridised to CGI microarrays. MAP fractionation of sperm DNA, repeatedly failed to yield sufficient material for microarray hybridisation under the equivalent experimental conditions applied to the somatic tissues (see methods). Despite the lack of quantitative data this observation empirically confirmed the hypomethylation of germ cell DNA relative to somatic tissues (Eckhardt et al., 2006; Kitamura et al., 2007; Weber et al., 2007). Signal data from MAP array hybridisations of the four somatic tissues (including blood), were extrapolated to give an enrichment profile for each (Fig. 4.3-1a). This result indicated that Brain had the smallest proportion of methylated CGIs (5.7%) with Muscle showing the most (8.3%). Due to the variability in methylation profiles between tissues a total proportion of 11.6% (1,657) of CGIs was identified as being methylated in one or more of the tissues tested.

To refine the comparison, CGIs with an M value in excess of 1.5 and a differential of ≥ 0.75 were considered as differentially methylated between the tissues. The result indicated the presence of a ubiquitously methylated set of islands representing 2.8% (408) of all CGIs (Fig. 4.3-1b and Table 4.3-1). Approximately 5% (711) of all CGIs were methylated in one or more but not all of the tested tissues, of which more than half (403) were methylated in one tissue alone (Fig. 4.3-1b; Table 4.3-1; Dataset 3). A proportion of the islands could not be categoriesed due to the thresholds applied (see methods for details).

The CGIs set was shown to colocalise with the transcriptional start sites of approximately 50% of all protein coding genes. Promoter-CGI methylation is known to correlate with transcriptional repression, and so gene localisation was compared for all CGIs and those found to be methylated in at least one of the tissues investigated. Islands were categoriesed into four gene association classes. Promoter associated (±1.5kb of an annotated TSS), 3' associated (±1.5kb of the most 3' terminus of the last exon), intragenic (associated with an open reading frame (ORF) but not falling into the former two categories) and Intergenic (not localising to within ±1.5kb of an annotated gene). Interestingly, methylated CGIs are relatively enriched at sites distal to the TSS. This is illustrated by the fact that 8% of all arrayed promoter CGIs are methylated relative to 21.7% of those associated with the extreme

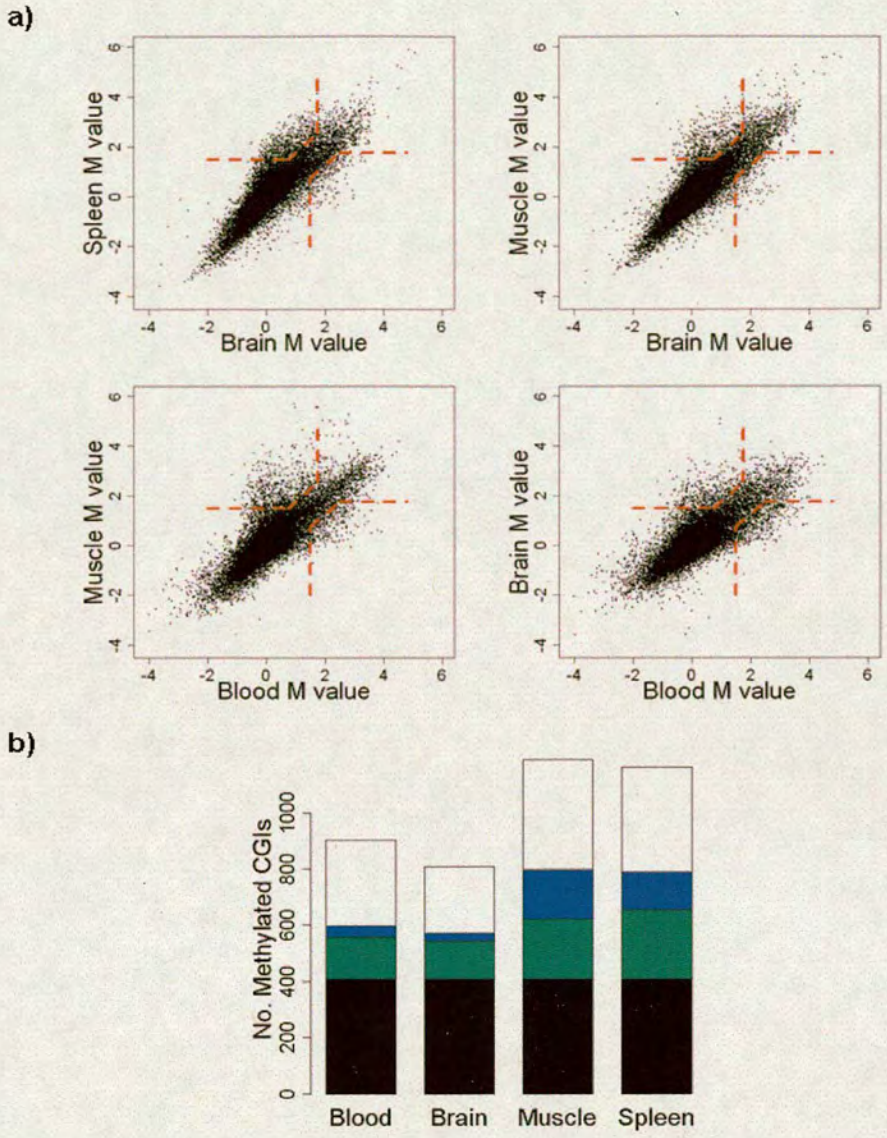


Figure 4.3-1. MAP array identification of tissue specific CGI methylation.
(a) Plots indicating pairwise comparisons between MAP CGI probe hybridisations prepared from blood, brain, muscle and spleen DNA. Dashed red lines indicate the threshold values applied to determine differential methylation. **(b)** CGI methylation events were subdivided into four different classes depending on the tissue specificity across the four tissues. The following categories are represented: CGIs methylated in all tested tissues (black); CGIs methylated in more than one tissue tested but not all (green); CGIs methylated in one tissue only (blue); CGIs methylated in one tissue tested but unclassified in other tissues (white).

3' terminus (Table 4.3-2). Indeed all classes of CGIs are at least twice as likely to be methylated as those proximal to promoter regions. This pattern is also observed for the subclass of differentially methylated islands (Table 4.3-2).

Table 4.3-1. CGI Methylation in Human Tissues

	Tissue Tested			
Methylation Status	**Blood**	**Brain**	**Muscle**	**Spleen**
Methylated in All	408	408	408	408
Differentially methylated (multiple[a])	149	135	214	247
Differentially methylated (single[a])	50	35	178	140
Unclassified methylation	303	237	392	381
Total	910	815	1192	1176
CGIs	14318	14318	14318	14318
Percentage Methylated	6.4	5.7	8.3	8.2

[a]Indicates the number of tissues containing a specific methylated CGI.

Table 4.3-2. Methylated CGI location relative to protein coding genes

CGI Gene association	All CGIs	Methylated	Methylated (%)	Differentially Methylated	Differentially Methylated (%)
All CGIs	14318	1657	11.6	711	5
Genes	11399	1225	10.8	514	4.5
5'	7926	636	78.0	256	3.2
3'	765	166	21.7	77	10.1
Intragenic	3478	536	15.4	230	6.6
Intergenic	2863	435	15.2	203	7.1

To confirm these results a panel of seven CGIs identified as differentially methylated between the tissues tested, were investigated using bisulfite genomic sequencing (Fig. 4.3-2 a-d and Fig. 4.3-6). In each case, microarray prediction was corroborated by the bisulfite data. CGI I1878 is distal to any annotated gene (>±1.5kb) and was found to be methylated exclusively in muscle and brain (Fig. 4.3-2a). The *SEC31B* gene transcribes a protein product which has been implicated in vesicular trafficking. Bisulfite analysis identified that a promoter CGI (I2985) which associates with the transcription start site of *SEC31B*, is compositely methylated in blood and spleen (Fig. 4.3-2b). CGIs I13406 (Fig. 4.3-2c) and I12175 (Fig. 4.3-6a) are methylated specifically in muscle and overlap the predicted gene *67313* and the 3' ends of *OSR1* respectively. I11878 locates to the 5' end of *ZNF577* and is only methylated in spleen (Fig. 4.3-2d). Interestingly two islands associated with the *PAX6* locus, namely I3654 and I3660, are both differentially methylated between the five tissues (Fig. 4.3-6b). CGI I3654, an island associated with the promoter region of an internal annotated *PAX6* transcript (*Q59GD2*), was previously shown to contain methylated CpG sites, and was found to be specifically methylated in brain (Nguyen et al., 2001); Fig. 4.3-6b). The second island is approximately 2kb upstream of the outer most annotated transcription start site and is heavily methylated in brain and muscle but devoid of

methylation in the other three tissues (Fig. 4.3-6b). Only one of the seven bisulfite-treated candidate CGIs directly overlapped an annotated TSS (Fig. 4.3-2b) consistent with the finding that methylated islands tend to occur outwith gene promoter regions.

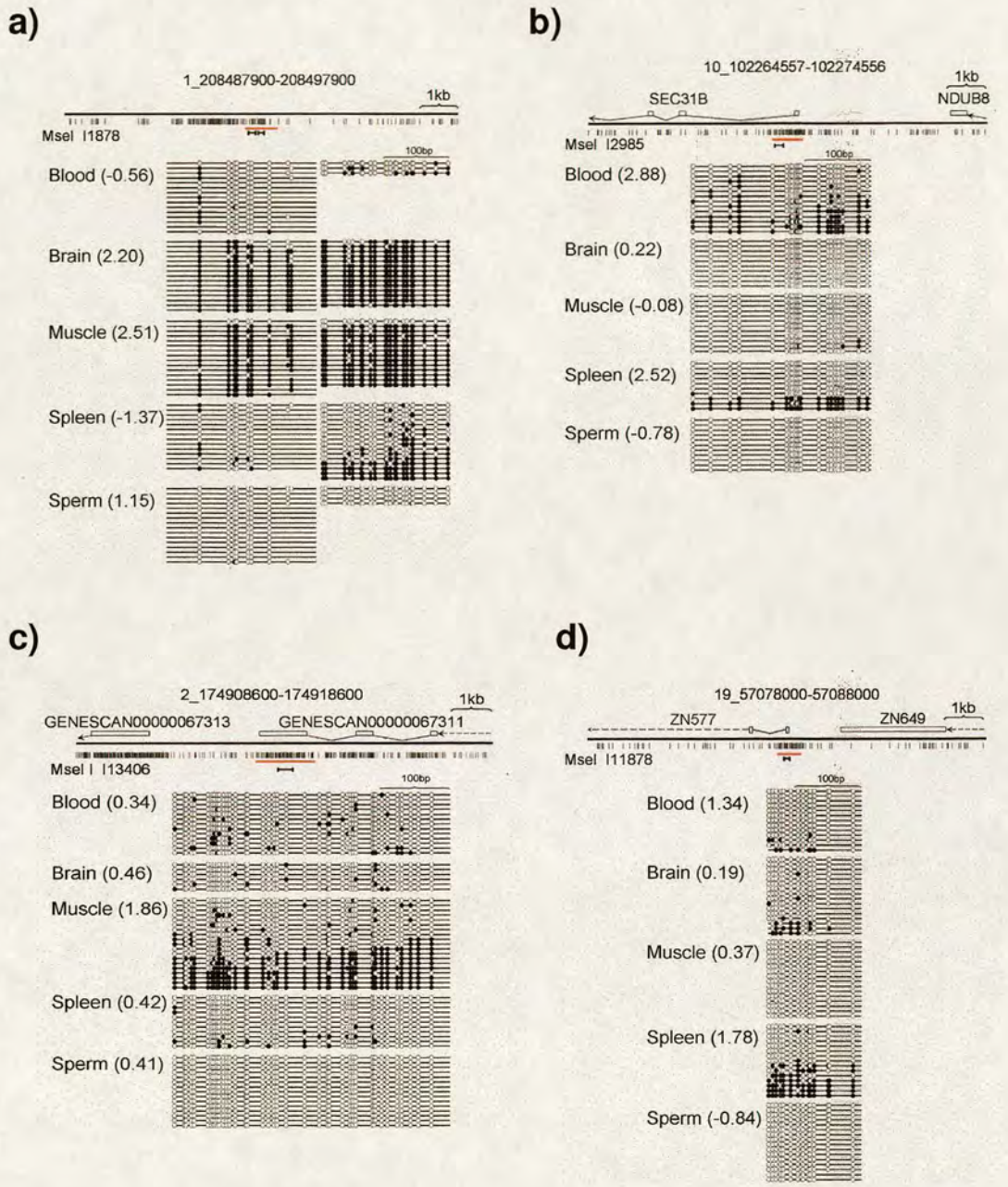


Figure 4.3-2. Confirmation of tissue-specific differential CGI methylation.
(**a-d**) Four CGI regions identified as being differentially methylated were assessed for DNA methylation by bisulfite genomic sequencing. *M values* are indicated for each tissue in parenthesis and the regions are illustrated as for Fig. 4.2-9b and c and annotated based on genomic build NCBI36.

4.3.2 Characterisation of Methylated CGIs

Previous studies have indicated that islands which are somatically methylated have a relatively reduced CpG density with respect to those which are not methylated (Weber et al., 2007). The assumption is that CpG methylation is concurrent with spontaneous methyl-cytosine to thymine transitions due to deamination. Consequently this would account for the relatively reduced density of the dinucleotide within these islands. In order to ascertain whether the methylated islands identified in this study had an altered sequence composition, the CpG density and G+C content of all CGIs found to be methylated (n=1,657) were compared with those that were not (n=12,661). This comparison identified a small but significant (Wilcoxen rank sum test: *p-value*: $1.022e^{-11}$) reduction in the CpG density of the methylated islands with respect to those not identified (median CpG[o/e] of 0.75 and 0.77 respectively; Fig. 4.3-3a). Despite the significant reduction in CpG density, the G+C content was slightly elevated in the methylated island set although the difference between the two sequence populations was again found to be small but significant (approximately 0.5% G+C content; Wilcoxen rank sum test: *p-value*: $<1.0\ e^{-3}$). Median sequence characteristics for all CGIs identified as methylated and those that were not are illustrated in Fig. 4.3-3a (blue and yellow asterisks respectively; see figure legend). The biological significance, of these small differences in sequence composition is unclear and may merely represent preferential localisation within the genome (this will be investigated in the next section). The fact that these methylated islands are less depleted for CpGs than has previously been reported may simply represent the different sets of sequences investigated in the two studies (Weber et al., 2007).

Using a panel of CGIs that were previously identified as being methylated by conventional methods, Bock and colleagues developed a prediction tool for *in Silico* identification of endogenous methylated CGIs (Bock et al., 2006; Yamada et al., 2004). After investigating a range of context specific sequence characteristics, they discovered that the set of methylated islands showed positive correlation with specific simple sequence patterns (CACC / GGTG and TGTG / CACA), repeat sequences (tandem repeats) and DNA structures. The authors stated that there was no single "DNA sequence code" which effectively described the methylation status of a CGI sequence (Bock et al., 2006). However, despite the subtle contribution of each of the afore mentioned sequence parameters, their combinatorial effect allowed relatively accurate prediction when applied to an independent set of methylated CGIs (Bock et al., 2006; Eckhardt et al., 2004).

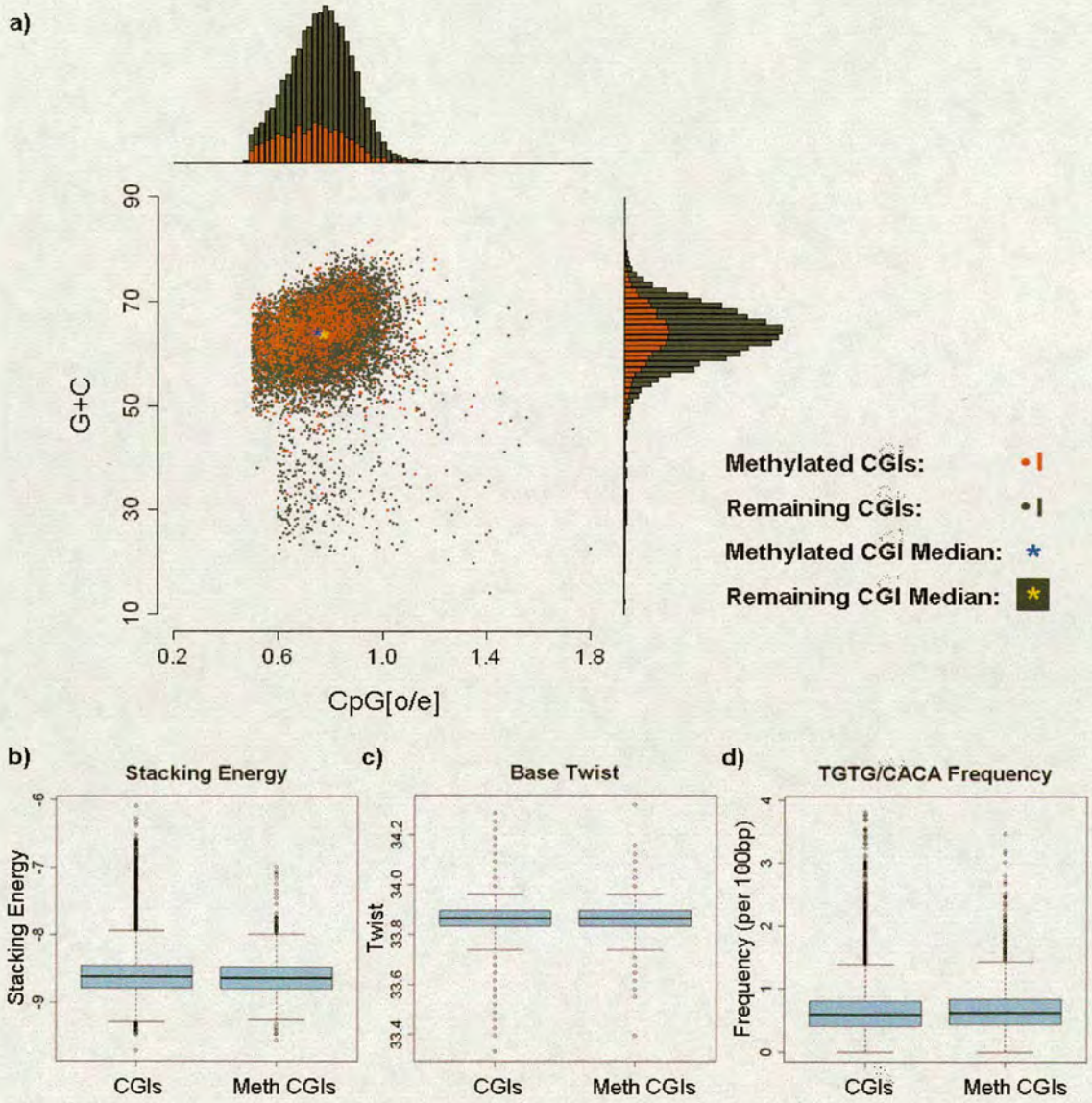


Figure 4.3-3. Comparison of physical properties between methylated and nonmethylated CGIs.

(a) Plot depicting the $CpG_{[o/e]}$ vs. G+C sequence compositions of CGIs identified as being methylated (red spots or histogram bars) or not (dark grey spots or histogram bars) and the corresponding histogram plots. These indicate a small but significant difference in the median sequence compositions of the two CGI sets (medians denoted by blue and yellow asterisks as per the figure key). Boxplots of stacking energy **(b)**, base twist **(c)** and TGTG/CACA frequency **(d)** comparing islands that were identified as being methylated (Meth CGIs) and those that were not (CGIs). Boxplots are depicted as for Fig. 3.1-1.

Given the concordance between these sequence characteristics and CGI methylation, we compared the attributes of the two CGI sets with respect to stacking energy, base twist[xxx],

[xxx] DNA structural parameters (twist and stacking energy) were determined using the EMBOSS package Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet *16*, 276-277.).

simple repeats (TGTG/CACA) and repetitive elements. Methylated CGIs show a small but significant increase in stacking energy relative to all CGIs (Wilcoxen rank sum test[xxxi]: *p-value* < 0.001; Fig. 4.3-3b). In contrast we found no significant difference in the base twist of methylated CGIs (Fig. 4.3-3c). TGTG/CACA specific repeats were found to be significantly enriched in methylated CGIs (Wilcoxen rank sum test: *p-value* < 0.001; Fig. 4.3-3d). In contrast to what has been previously describe, all repetitive elements (as outlined in Repbase (Jurka et al., 2005)) were found to be marginally depleted in methylated CGIs (Wilcoxen rank sum test: *p-value* < 0.01).

Despite the apparent concordance between these sequence characteristics, not all match previous predictions, and those which do show modest differences between the CGI sets. Whilst it is possible that these features are characteristic of methylated CGIs and can distinguish them from CGIs in general it is equally possible that these represent a by-product of their genomic locale? To investigate this possibility the location of all methylated CGIs was plotted against the 22 human autosomes (Fig. 4.3-4a). However this alone was insufficient to determine if specific regions of the genome were more enriched as all CGIs are discretely localised throughout the genome (as discussed in the previous chapter). In order to address this problem an average autosomal profile of CGI and methyl-CGI distribution was determined. Each of the 44 chromosome arms were subdivided into windows of 5% of the total arm length. The CGI and methylated CGI content of each window was measured in turn with a periodicity of 1% until the centromere was reached (96 windows). This process was repeated for each chromosomal arm and the values summed. The distributions of methylated and nonmethylated CGIs were then plotted against an averaged chromosomal arm (Fig. 4.3-4b). To refine this further, the observed number of methylated islands per window was divided by that expected if methylated islands represented a random subpopulation of all CGIs[xxxii]. The averaged Methylated $CGI_{[o/e]}$ autosomal arm schematic indicated that methylated CGIs were relatively enriched in subtelomeric and pericentromeric regions (Fig. 4.3-4c). The fact that there are relatively few islands adjacent to the centromeres suggests that the apparent enrichment of methylated CGIs could well be an artifact of sample size (Fig. 4.3-4b and c). However, as there is high density of CGIs found in subtelomeric regions, it is likely that the two-fold

[xxxi] Data distributions were tested using Shapiro-Wilk test of normality. Non parametric significance values were determined using the Wilcoxon rank sum test between the Methylated and remaining sets of CGIs (n=1657 and 10661 respectively).

[xxxii] Expected island count was calculated by dividing the sum of all methylated CGI for each window by the equivalent sum of all CGIs. The count of islands was then multiplied by this fraction to give the expected number fro each window.

overrepresentation of methylated islands indicates a genuine spatial enrichment within the genome.

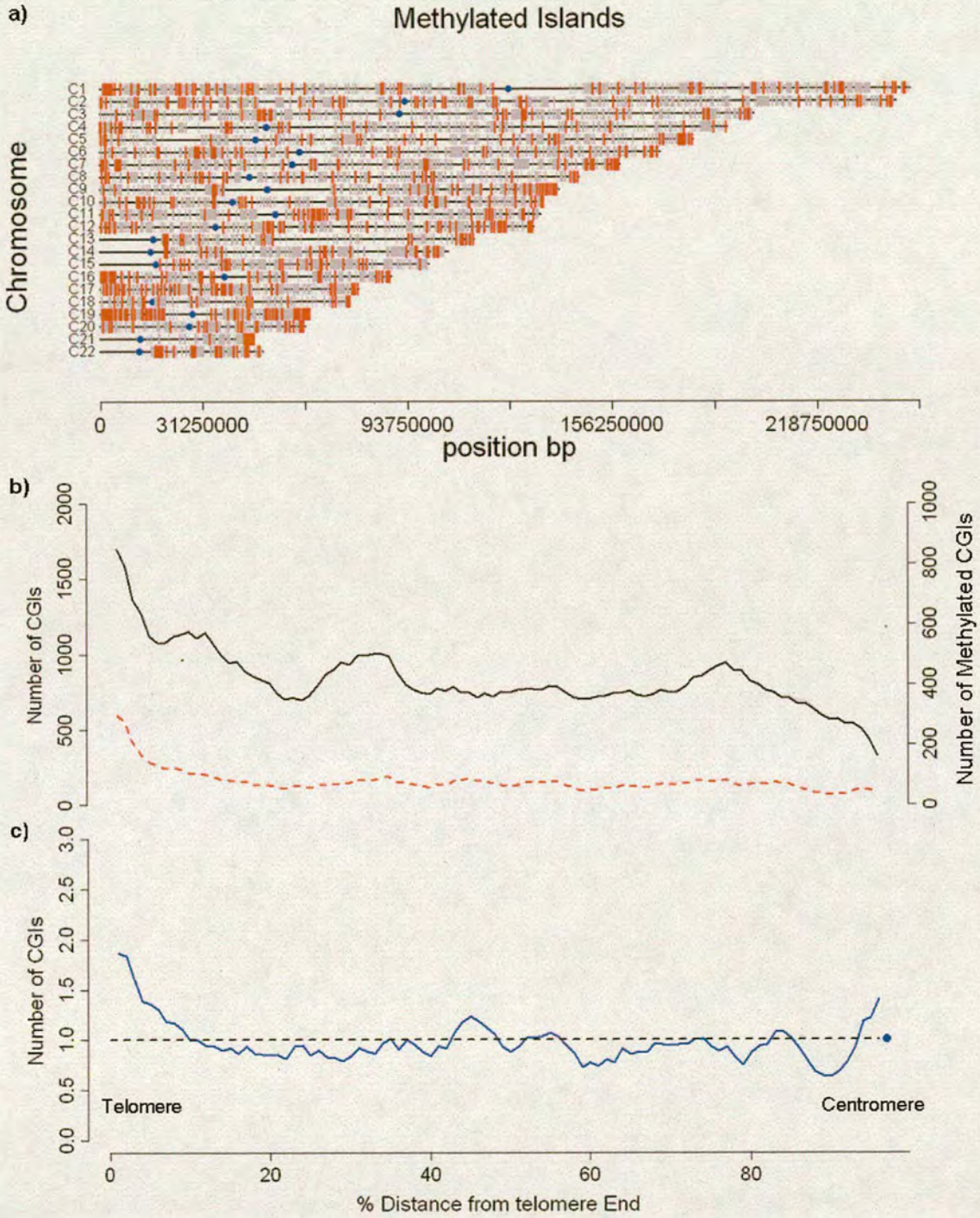


Figure 4.3-4. The distribution of all methylated CGIs across the human autosomes.
(a) The distribution of cloned CGIs (grey strokes) on human chromosomes overlaid with all MAP array identified methylated CGIs (Red Strokes). Chromosome numbers are indicated to the left of each chromosome and centromeres are denoted by blue dots. **(b)** Plots depicting the CGI (black line; primary Y axis) and methylated CGI (dashed red line; secondary Y axis) distribution across all autosomal arms. **(c)** Plot indicating the observed / expected [o/e] methylated CGI density across an average autosomal arm. An [o/e] ratio of 1 (dashed line) and centromere position (blue point) are indicated. Plots **(b and c)** represent a window size and sliding resolution of 5% and 1% of autosomal arm length respectively.

This finding is interesting as it suggests that the subtelomeric domains are enriched for methylated CGIs. Furthermore, it is consistent with a previous study which cloned methylated CpG rich methylated DNA fragments derived from telomere-proximal regions (Brock et al., 1999). These regions are also enriched for a class of genes previously identified as being associated with methylated CGIs. Odorant receptors are known to be methylated in somatic tissues and to be enriched in subtelomeric regions (Riethman et al., 2004; Weber et al., 2007). This finding is of relevance to the mechanisms by which certain CGIs become methylated and the organization of the mammalian genome. However it does not confirm the hypothesis that sequence disparities result from this spatial genomic arrangement.

To address this possibility, all CGIs located at the telomere ends[xxxiii] were extracted and their sequence characteristics compared to the remaining CGIs. CpG density was unaffected by this subdivision whereas the G+C composition of the telomeric CGIs (median = 66%) was significantly elevated relative to the remainder of the islands consistent with previous observations (median = 64%; Wilcoxen Rank sum test: *p value*: $2.2e^{-16}$; (Riethman et al., 2004)). Although the biological significance of the altered sequence characteristics of methylated CGIs remains unclear this provides some evidence for spatial clustering within the genome.

4.3.3 Composite Methylation of CGIs

All methylated CGIs thus far characterised by bisulfite genome sequencing have displayed a binary methylation pattern composing of nonmethylated and heavily methylated DNA strands (Fig. 4.2-9c; Fig. 4.3-2a-d; Fig. 4.3-5b and c; Fig. 4.3-6a and b). This is illustrated when the average percentage of methylated CpG sites present on each DNA molecule is plotted for each of the methylated islands assayed (Fig. 4.3-5a). This analysis identified a bimodal profile of methylation, indicating a 'black and white' distribution. A study dissecting the methylation profile of specific DNA sequences on human chromosomes 6, 20 and 22 identified a similar phenomenon (Eckhardt et al., 2006). This finding solves the apparent paradox arising from the ability to identify methylated islands in blood using an array of CpG rich nonmethylated sequences.

[xxxiii] Telomere proximal regions were defined as the outer most 10% window of each autosomal arm.

It is possible to suggest several biological explanations for these methylation profiles. At the highest level, these composite patterns could simply represent individual methylation polymorphisms within the pooled DNA. To address this possibility, compositely methylated CGI, I5134 was bisulfite sequenced from the muscle DNA of each of the three pooled individuals. Strikingly this confirmed that individual C had significantly higher levels of methylation at this particular CGI (Fig. 4.3-5b). Furthermore, MAP array identified I24201, a CGI associated with the *ENDOG* gene, as being methylated specifically in spleen. Interestingly a previous study identified this island as being compositely methylated in normal lung tissue in two of eight individuals (Shiraishi et al., 1999).

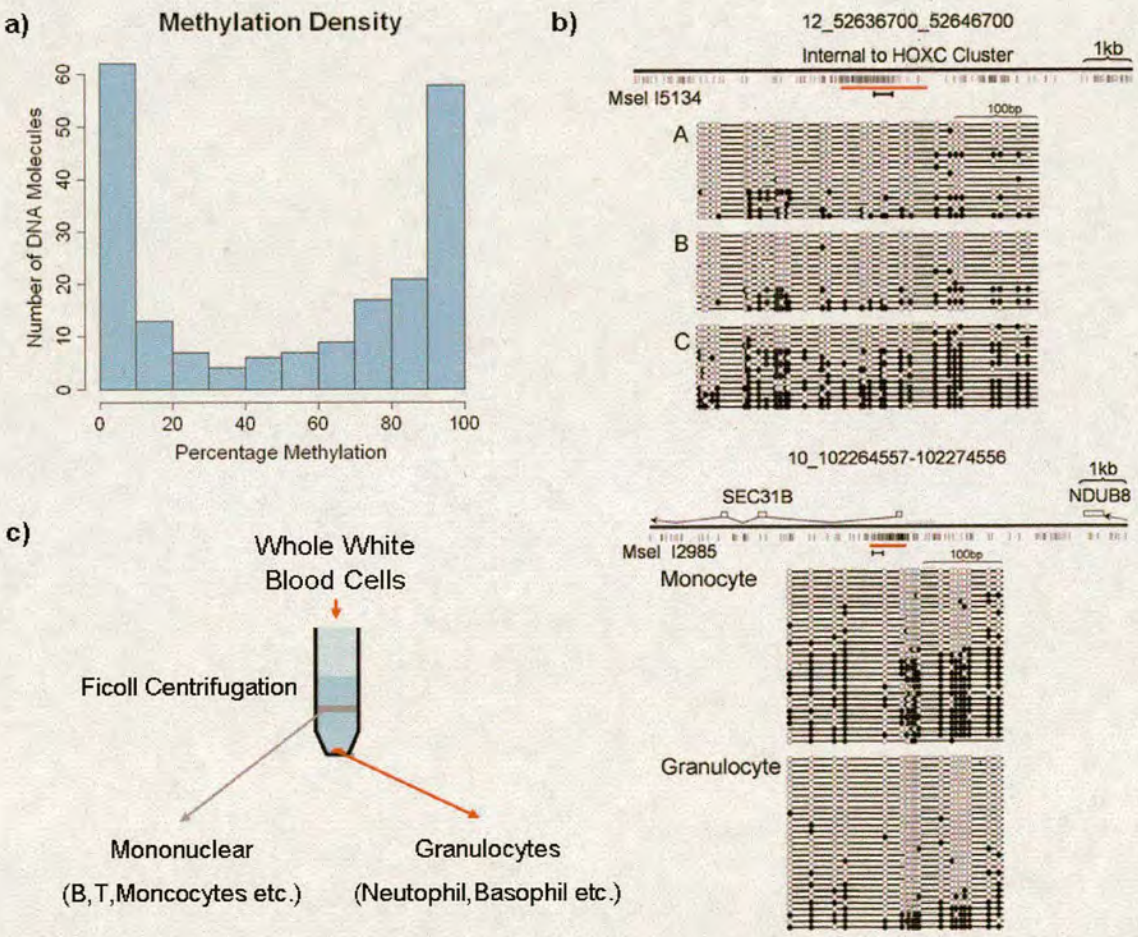


Figure 4.3-5. Composite DNA methylation patterns of CGI sequences.
(a) A histogram depicting the percentage of methylated CpG sites per DNA strand identifies a bimodal distribution of essentially nonmethylated and heavily methylated DNA molecules. **(b)** Bisulfite sequencing identifies individual-specific CGI methylation internal to the HOXC cluster in muscle DNA. **(c)** Bisulfite sequence analysis of DNA derived from mononucleocyte and granulocyte white blood cell preparations identifies cell type specific enrichment of DNA methylation at the *SEC31B* promoter. Bisulfite genomic sequencing results (**b** and **d**) are diagrammed as in Fig. 4.2.9.

Alternatively, cell specific CGI methylation patterns could account for some of the composite methylation identified. The methylation pattern of CGI (I2985), previously shown to be compositely methylated in blood and spleen DNA (Fig. 4.3-2b), was examined in fractionated mononucleocyte and granulocyte cells. Interestingly this CGI was found to be more heavily methylated in mononucleocytes than granulocytes prepared from the same individuals (Fig. 4.3-5c). This indicates that at least some of the CGI methylation is cell type specific. Analysis of the DNA methylation pattern at a CGI associated with the murine *PAX6* locus determined that like it's human counterpart, it was also compositely methylated. Further analysis determined that the methylated and nonmethylated DNA molecules were not resolved when neuronal and glial DNA sequences were characterised (RI and Peter Skene unpublished observations).

Other instances where composite methylation could not be resolved as for these examples, suggests an alternative origin for this distinct methylation pattern. Allele specific methylation is generally well characterised in the context of imprinted DMRs but parent independent allelic methylation has also been illustrated in a few cases (Yamada et al., 2004). In the absence of informative single nucleotide polymorphism (SNP) data and pedigree information it was impossible to relate the methylation profiles directly to specific alleles. However in an attempt to address this possibility, all blood[xxxiv] methylated CGIs were compared to a list of genes identified as being monoallelically expressed in a global screen (Gimelbrant et al., 2007). Of the 228 unique protein coding genes which mapped to arrayed CGIs, only 17 were found to be methylated by MAP array analysis (Table 4.3-3). This finding suggests that monoallelic methylation is not a major contributor to the observed composite pattern. However it is impossible to conclude the concordance between the methylated islands and monoallelic expression without matched allelic expression data from the same biological samples for which the DNA methylation patterns were derived.

4.3.4 Differential CGI methylation and developmental gene loci

Differentially methylated CGIs identified by MAP array showed elevated spatial localisation to developmental genes and gene loci (Fig. 4.3-6). In order to investigate this further, ontology terms for gene associated CGIs were compared with those identified as differentially methylated (Table 4.3-4 and Dataset 3). Interestingly, genes involved in developmental processes such a neurogenesis; segmentatation specification and mesoderm

[xxxiv] Blood methylation was chosen for two reasons. The first is that due to the nature of the library preparation the likelihood of blood methylated islands being compositely methylated is high. Secondly monoallelic expression gene expression data was determined in peripheral blood mononucleocyte and lymphocyte cells.

development were significantly enriched for this group of islands (Table 4.3-4). Furthermore, the Homeobox class of developmental transcription factors was more then 3 fold overrepresented than would be expected by proportional representation (Table 4.3-4). Of the genes characterised by bisulfite genomic sequencing, *HOXC* is involved in body

Table 4.3-3. Blood Methylated CGIs that associate with Monoallelically expressed genes.

Mono Allelic CGI Gene		Methylated CGI Information				
Gene ID[a]	Gene Name[b]	CGI ID[c]	Gene Overlap	Chr	Start[d]	Stop[d]
ENSG00000196581	AJAP1	I68	intragenic	1	4671134	4673025
ENSG00000136010	ALDH1L2	I5530	5 prime	12	104002162	104003256
ENSG00000168675	C18orf1	I10355	intragenic	18	13631490	13633034
ENSG00000050767	COL23A1	I19508	intragenic	5	177935057	177936975
ENSG00000113758	DBN1	I19491	intragenic	5	176832031	176833615
ENSG00000183495	EP400	I5846	5 prime	12	130999923	131001992
ENSG00000115641	FHL2	I12903	intragenic	2	105381239	105382810
ENSG00000152990	GPR125	I17114	5 prime	4	22125920	22127617
ENSG00000131398	KCNC3	I11851	intragenic	19	55522437	55524577
ENSG00000104044	OCA2	I7431	intragenic	15	25821255	25822469
ENSG00000178184	PARD6G	I10831	intragenic	18	76018459	76019662
ENSG00000139832	RAB20	I6494	5 prime	13	110011166	110012522
ENSG00000080293	SCTR	I13000	5 prime	2	119998052	119999037
ENSG00000198648	STK39	I13334	5 prime	2	168811817	168812986
ENSG00000187079	TEAD1	I3517	intragenic	11	12843196	12843336
ENSG00000142185	TRPM2	I14898	5 prime	21	44594267	44596251
ENSG00000196793	ZNF239	I2518	5 prime	10	43389437	43390252

[a] Gene IDs are based on ENSEMBL transcript annotation.
[b] Gene Names: HGNC names.
[c] CGI identifiers correspond to microarray annotation and the IDs mapped to ENSEMBL.
[d] CGI Start and Stop sequence locations are based on NCBI v.36.

pattern specification, *PAX6* is involved in eye development and *OSR1* is related to a *Drosophila* gene encoding a protein involved in gut development (Fig. 4.3-5b and Fig. 4.3-6a and b). The MAP array results for the extended *PAX6* and *HOXC* loci were mapped to determine the variability in methylation across these developmental regions (Fig. 4.3-6c and d). The *PAX6* loci identified the 2 CGI regions which were differentially methylated, in brain and muscle as confirmed by bisulfite genomic sequencing (Fig. 4.3-6b and c). The 150kb *HOXC* locus is composed of 19 CGIs of which eight were found to be differentially methylated in blood, muscle and spleen with none being methylated in brain (Fig. 4.3-6d). The observed variability is interesting as it suggests that expanding the panel of tissues would identify additional differentially methylated islands.

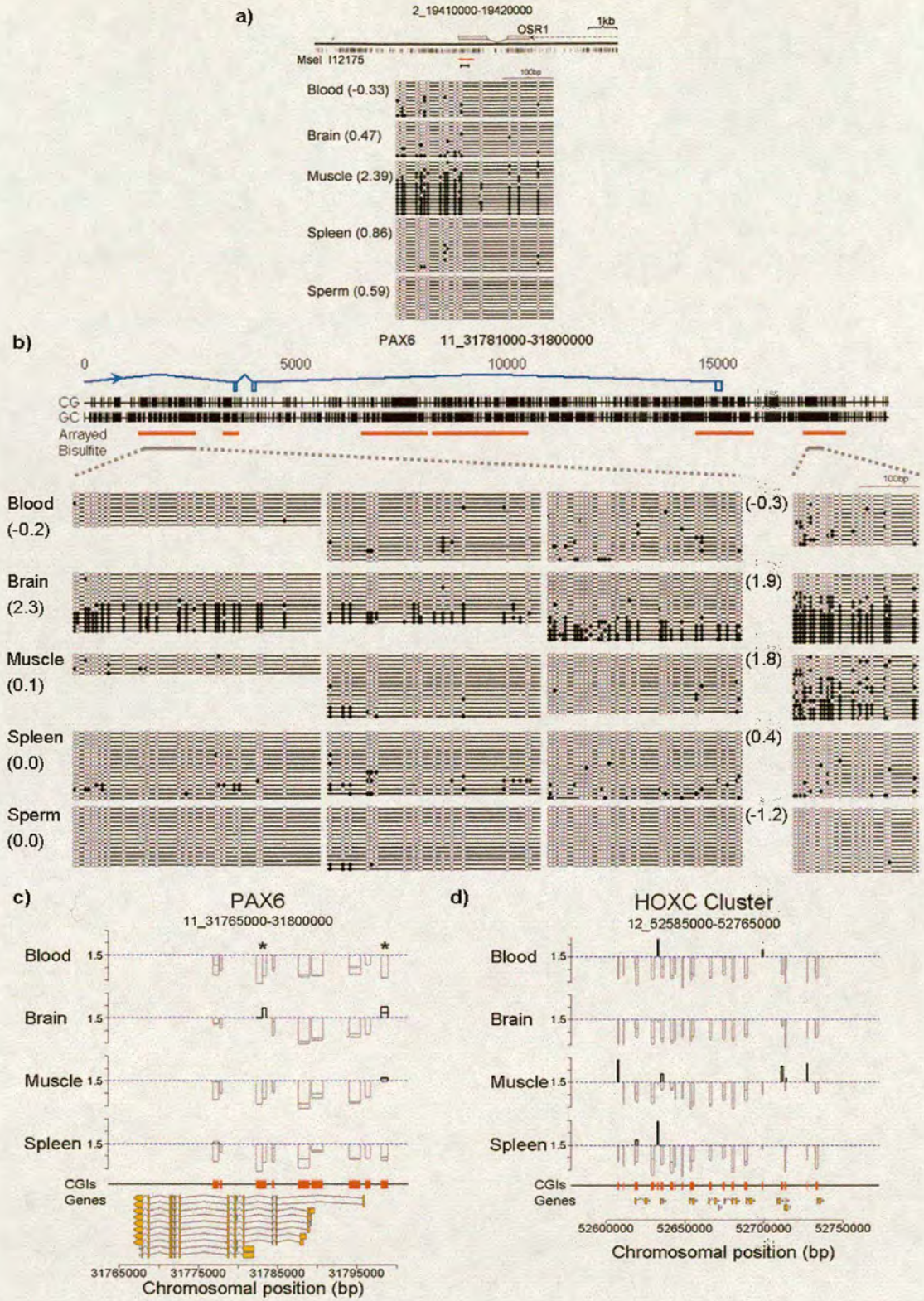


Figure 4.3-6. Enrichment of CGI methylation at developmental gene loci.
Bisulfite genomic sequencing confirmed tissue-specific CGI methylation associated with the developmental genes OSR1 (**a**) and PAX6 (**b**). MAP array determined M values are indicated in parenthesis for each tissue. The OSR1 (a) and PAX6 (**b**) bisulfite results are diagrammed as for Fig. 4.2-9. Arrayed CGI amplicons located across the PAX6 region (**b**) are indicated (red bars) in addition to all CpG (CG) and GpC (CG) dinucleotide positions (vertical black strokes). The exonic structure of

the major PAX6 isoform (blue boxes), transcriptional direction (arrowed) and amplicons assayed for methylation depicted (solid grey bars) are indicated. (c and d) Multiple CGIs (red boxes) span the PAX6 (**c**) and HOXC (**d**) gene loci. MAP-CGI array profiles for blood, brain, muscle and spleen identify tissue-specific CGI methylation (vertical black bars extending above M = 1.5). Grey bars extending downwards below M = 1.5 (broken blue line) represent non-methylated CGIs. The regions of PAX6 analysed by bisulfite genomic sequencing (see b) are indicated (asterisks in c). Tick marks on the y axis are spaced at intervals of 1 M value unit. Coding sequences are diagrammed as yellow bars.

The observation that tissues presented distinct patterns of CGI methylation and that these often corresponded to genes with developmental functions posited the possibility that methylated CGIs function in tissue specific gene regulation. To address this possibility, gene expression data was mined for all CGIs associated with a gene TSS found to be methylated in a single tissue. Promoter associated CGIs were selected as all known cases of methylation at such sequences corresponds to gene repression. Alternatively, the transcriptional affect of gene body or 3' CGI methylation remains unclear. Expression data mining was carried out using the GNF Symatlas (http://symatlas.gnf.org/SymAtlas/) web based browser tool (Su et al., 2002). The analysis indicated no obvious accord between the methylation status of these islands and the expression of the associated genes. Despite this fact, methylated islands always showed an expression value equal to or less than the median value for all tissues examine. However, often the genes were silent in the majority of all tissues irrespective of methylation status. This is illustrated for GPNMB, which associates with CGI I21166 identified by MAP array as being methylated in skeletal muscle alone. The gene appears to be silent in whole blood, muscle and whole brain despite being methylated in blood alone (Fig. 4.3-7a). Expression analysis of SEC31B gene identified as being more heavily methylated in mononucleate blood cells (Fig. 4.3-5c) indicated that it was silent in whole blood but not in CD4+ and CD8+ T cells (Fig. 4.3-7b). This is surprising as these cells should be present in the mononuclear blood cell fraction and therefore be expected to be expressed at a reduced level. In order to address this further, the specific expression of the SEC31B gene was determined from RNA prepared form the mononucleocyte and granulocyte cell preparations used to generate the bisulfite methylation profiles (Fig. 4.3-5c). Contrary to the prediction there was little expression difference between the two cell populations when normalised to GAPDH expression irrespective of promoter methylation status (Fig. 4.3-7c). Consistent with much of the mined data, expression levels were low for all tissues, independent of promoter CGI methylation. This result suggests promoter CGI methylation is possibly a consequence of transcriptional quiescence rather than a cause. Alternatively, similar to the situation observed for X-inactivation, promoter CGI methylation may not represent an initiating event in transcriptional repression, rather than stable propagator of the repressed state.

To address these uncertainties, future studies will have to map global tissue specific CGI methylation events to gene expression levels in matched biological samples. Furthermore, individual variation will have to be accounted for should polymorphic methylation play a role in individual phenotypic variation.

Table 4.3-4. Developmental gene categories are associated with differentially methylated CGIs.

Biological Process[a]	All genes[b] (n=9542)	Methylated (observed n=490)	Methylated (expected)	p.value[c]
Developmental processes	1187	112	60.95	4.00E-09
mRNA transcription reg'n	847	85	43.5	4.61E-07
Ectoderm development	415	51	21.31	1.94E-06
mRNA transcription	1093	94	56.13	6.42E-05
Neurogenesis	383	45	19.67	6.86E-05
Segment specification	71	14	3.65	3.77E-03
Mesoderm development	326	35	16.74	6.61E-03

Molecular Function	All genes[b] (n=9542)	Methylated (observed n=490)	Methylated (expected)	p.value[c]
Homeobox TF	153	37	7.86	3.16E-12
Transcription factor	1131	99	58.08	2.54E-06
Other DNA-binding protein	180	26	9.24	5.66E-04

[a] GO terms were determined using the web based Panther classification system (Thomas et al., 2003).
[b] All genes associated with an arrayed CGI
[c] Significance is corrected for multiple testing (Bonferroni adjustment).

4.4 Discussion

In recent years, much investigation pertaining to mammalian DNA methylation has focused on human neoplasias. This has resulted in a relative paucity of information regarding the distribution of methylation in normal human cells. For CGIs, the characteristic lack of methylation, an important empirical criterion of this discrete genomic fraction, was considered to be the general rule. Obvious exceptions to this included methylated CGIs on the inactive X chromosome, allelic CGI methylation at imprinted loci and CGIs associated with certain germ line specific genes. These unique, functionally important, exceptions to the dogma were followed by increasing evidence suggesting that somatic CGI methylation may be more prevalent than previously indicated. This chapter described the characterisation of the CGI methylomes in a panel of human tissues.

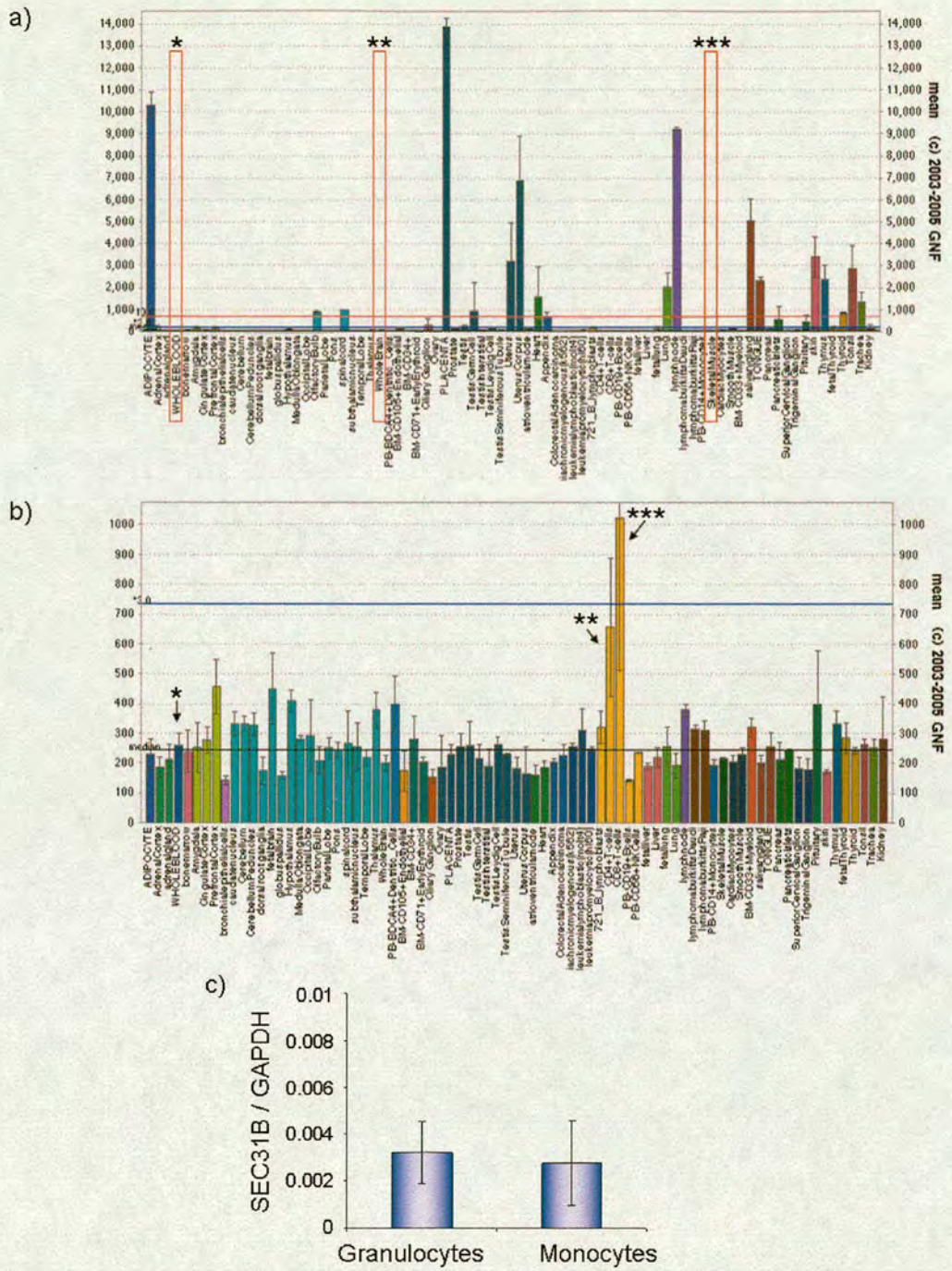


Figure 4.3-7. Gene Expression and differential CGI methylation

Expression profiles for *GPNMB* (**a**) and *SEC31B* (**b**) produced by the visualization of array determined transcriptome data using the GNF Symatlas. (**a**) The *GPNMB* expression values for whole Blood (*), whole brain (**) and skeletal muscle (***) are highlighted (red boxes) indicating relatively low expression levels in comparison to the panel of tissues assayed. (**b**) The expression of the *SEC31B* gene is low in whole blood (*) but elevated in CD4+ (**) and CD8+ T cells (***). (**a** and **b**) Error bars represent SD values between multiple hybridisations and the median (grey line) and 3xmedian (blue line) are depicted for each gene. Images were generated and adapted from those generated by the Symatlas visualization tool (Su et al., 2002). (**c**) A plot of *GAPDH* normalised *SEC31B* expression indicates equivalent low level expression in Mononucleate and Granulocyte cell fractionations from whole white blood cells. Expression values depicted are averaged from two biological experiments repeated in triplicate, the variance for which is denoted by the error bars (SD).

4.4.1 MAP array development and optimisation

MAP has previously been shown to be a powerful method for the enrichment of methylated CpG rich DNA sequences from bulk genomic DNA (Brock et al., 1999; Cross et al., 1994; Selker et al., 2003; Zhang et al., 2006). Bacterial induction of the published MBD construct indicated that it expressed poorly and was almost exclusively restricted to insoluble inclusion bodies. To circumvent this problem an alternative construct representing an equivalent portion of the human MeCP2 MBD domain was cloned in frame with a carboxy-terminal histidine tag. The fusion protein was expressed at much higher levels and was found to be predominantly soluble throughout lysis and purification. The efficacy of the MBD for application to MAP was assessed by fractionation of artificially methylated plasmid fragments. Calibration indicated that nonmethylated sequences eluted at approximately 0.75M NaCl concentration, whereas methylated sequences were retained until NaCl concentrations in excess of 0.85M. As for CAP, the affinity matrix had a preference for sequences bearing multiple CpG sites. This did not directly indicate elevated CpG density as a prerequisite for binding *per se*, but did suggest that short fragments derived from bulk genomic DNA would have insufficient CpGs to be retained by the affinity matrix. Relatively low resolution separation between nonspecific and specific DNA sequences was countered by the introduction of an optimized wash step at 0.75M NaCl. Blood DNA fractionation indicated that these conditions efficiently removed the majority of bulk genomic DNA, whilst candidate PCR analysis confirmed the retention of endogenous methylated CGIs. Buffering conditions applied were consistent with those previously published (Brock et al., 1999; Cross et al., 1994).

DNA from 4 human somatic tissues and sperm were fragmented and prepared using this optimised MAP protocol. High affinity fractions from each tissue were amplified, fluorescently labeled and hybridised to microarrays allowing the simultaneous screening of 60% of all human CGIs. Male and female MAP hybridisations identified that more than 40% of X linked CGIs were specifically methylated in female but not male DNA consistent with X inactivation. In addition to confirming the efficacy of the technique, this result indicated that sensitivity of MAP array was sufficient to identify monoallelic methylation.

A previous study identified a sequence preference for the MBD of MeCP2 for a run of A and T residues adjacent to the central CpG site (Klose et al., 2005). This presented the possibility that MAP, utilising this affinity construct, may inadvertently be biasing sequence selection. However on inspecting the MAP array results, there was no apparent enrichment for

particular sequence compositions. It is likely that DNA containing multiple CpG sites provided sufficiently high affinity ligands for nonbiased enrichment. This may however, indicate that future application of MAP array to the analysis of more CpG deficient portions of the genome will be biased. This could be avoided through application of the MBD2 MBD which has no apparent sequence preference outwith the core CpG dinucleotide. This was illustrated previously by the effective purification of methylated sequences by this domain (Gebhard et al., 2006a; Gebhard et al., 2006b; Klose et al., 2005).

4.4.2 Somatic CGI methylation

Insufficient sperm DNA was retrieved from MAP fractionations to label for microarray analysis. Consistent with previous findings, this observation provided empirical evidence that sperm CGIs are hypomethylated relative to those in somatic tissues (Brock et al., 1999; Weber et al., 2007). MAP array identified 7.8% of transcription start site associated CGIs as methylated in one or more of the tissues tested. This is significantly more than the 3% of promoter associated CGIs, identified in a related study (Weber et al., 2007). This disparity may have arisen, in part, from the inclusion of tissue specifically methylated islands missed in the latter investigation. Indeed 5% of all islands were identified as being differentially methylated in a subset of the tissues tested. Bisulfite genomic sequencing of 512 CGIs on chromosomes 6, 20 and 22 identified 9.2% of CGIs as hypermethylated (Eckhardt et al., 2006). This non-gene centric analysis of CGI methylation provided a similar result to the 11.6% identified here.

Sequence characterisation of CGIs suggested that those with high CpG density were somewhat impervious to DNA methylation (Eckhardt et al., 2006). Consistently, a small but significant reduction in CpG density was observed for the somatically methylated islands. Interestingly, G+C composition was also slightly elevated for the methylated CGIs, consistent with their observed enrichment in telomere proximal regions (Brock et al., 1999). Elevated G+C composition may therefore reflect a specific spatial distribution of methylated islands rather than a distinct sequence composition.

The regulatory potential for CGI methylation on gene expression, has led to the identification of sequence characteristics indicative of methylation (Bock et al., 2006). Simple repeats, sequence patterns and certain physical characteristics were identified as being elevated specifically in this subset of CGIs. Consistently, stacking energy and the simple repeat TGTG / CACA were found to be significantly enriched. However base twist

was unaffected and repetitive elements were somewhat depleted in methylated CGI sequences. Weber and colleagues reported that sequences intermediate in CpG density to that characteristic of CGIs and bulk genomic DNA were more frequently substrates for *de novo* methylation (Weber et al., 2007). On comparison with the novel CGI set employed here, it was apparent that 75% of these were not represented. This may be explained by the increased abundance of MseI cleavage sites, depleted CpG density or increased DNA methylation in this class of sequences. This would prevent sufficient retention of these sequences by the CAP and subsequently exclude them from the library. Consistent with this hypothesis, 22 intermediate CpG density sequences found to be methylated were absent from the CGI set. This suggests that the observation of altered sequence characteristics previously described may be somewhat confounded by the prerequisite removal of methylated and CpG deficient sequences from the CGI set. As such, the library will likely represent a somewhat unique class of islands relative to those identified by less biologically relevant means.

4.4.3 Composite CGI Methylation

Interestingly, bisulfite genomic sequencing of methylated CGIs uncovered a composite methylation pattern for almost all sequences tested. This bimodal distribution suggests that the DNA molecules were largely methylated or unmethylated with relatively little intermediate methylation. This distinct pattern may represent individual methylation polymorphisms resulting from the pooling of DNA samples. Indeed, detailed characterisation of a CGI associated within the *HOXC* cluster was considerably more methylated in one of the three individuals. Furthermore, a methylated island which associates with the *ENDOG* promoter was previously identified as being differentially methylated in a panel of lung samples (Shiraishi et al., 2002). An alternative possibility is that different cells types, composing the tissues investigated, bear distinct methylation patterns. Analysis of a *SEC31B* associated CGI, found to be compositely methylated in whole blood was more heavily methylated in mononucleocytes than granulocytes prepared from the same individuals. Finally, monoallelic methylation could also account for the observed methylation pattern. A study investigating the methylation status of CGIs on chromosomal arm 21q, identified both imprinted and parent independent monoallelic methylation (Yamada et al., 2004). Without SNP or familial linkage data this was impossible to test directly. However to address this, promoter associated CGIs were compared with a list of independently identified monoallelically expressed genes (Gimelbrant et al., 2007). CGIs methylated in blood were characterised, as the likelihood of identifying islands bearing this distinct pattern would be improved by the nature of the CGI library. This comparison

identified 17 genes that are candidates for allele specific methylation. Despite this relatively small overlap it is likely that this phenomenon will contribute, in part, to the observed composite methylation pattern characteristic of this subset of CGIs.

4.4.4 CGI methylation, gene association and expression

Promoter CGI methylation has previously been shown to correspond with transcriptional gene repression (Hansen and Gartler, 1990; Stein et al., 1982). Consistently, CGI genes which escape X-inaction have been shown to evade the *de novo* methylation characteristic of the inactive chromosome (Weber et al., 2005). MAP array confirmed that CGI genes which escape X inactivation, were significantly less methylated than their inactive counterparts (Carrel and Willard, 2005). This observation supports the notion that transcriptional inertia corresponds to promoter CGI methylation. As such we hypothesized that CGI genes which were differentially methylated would show a concordant reduction in gene expression. However, batch mining of tissue specific expression data yielded no obvious correlation between methylation and gene activity. It is possible that specific effects on gene expression were somewhat diluted by the composite methylation pattern observed. To address this possibility, *SEC31B* expression was assessed in mononucleocytes and granulocytes where the promoter CGI is differentially methylated. This candidate analysis identified equivalent low level expression in both cell populations. This finding was consistent with the observation that many of the differentially methylated islands were expressed at very low levels in all tissues irrespective of methylation state. No attempt was made to correlate internal CGI methylation and gene expression, as the affect of this phenomenon is not evident at this time. The significance of these observations will be discussed later but it is clear that future methylation studies will have to be associated with matched expression analysis.

As an alternative means of investigating the role of CGI methylation and gene function, all differentially methylated CGI genes were classified according to their gene ontologies. Interestingly, this indicated that differentially methylated CGIs associated with genes involved in developmental processes such as neurogenesis and ectoderm development. Furthermore, homeobox transcription factors involved in body segment specification were more than 3 fold over represented with 25% of all HOX cluster genes associating with a differentially methylated CGI sequence. The role of this methylation remains unclear as many of these islands are distal to promoter regions. It is tantalising to posit that these methylated islands may play a functional role by mediating the expression of an antisense

ncRNAs. Indeed, both the *Air* and *Tsix* ncRNA transcripts originate from CGIs and are involved in the regulation of the sense transcript (Panning and Jaenisch, 1996; Sleutels et al., 2002; Wutz et al., 1997). Furthermore, evidence for *HOX* cluster regulation by the tissue specific expression of the ncRNA *HOTAIR* further highlights the necessity for further investigation of these interesting developmental gene loci (Rinn et al., 2007).

Chapter 5: Discussion

Current technical limitations have prevented the characterisation of the entire human DNA methylome; a feat which has only been achieved for the small genomed plant *Arabidopsis thaliana* (Cokus et al., 2008; Zhang et al., 2006; Zilberman et al., 2007). Therefore, attention has been directed towards sequences of evident biological interest and more pragmatically, those which are tractable to molecular scrutiny in the lab. CGIs are conspicuous within the human genome as they have an atypically elevated CpG and G+C composition (Bird, 1987; Cooper et al., 1983; Gardiner-Garden and Frommer, 1987). They are generally hypomethylated and transcriptionally permissive in all tissues (Cooper et al., 1983; Tazi and Bird, 1990). CGIs associate with the promoters of more than half of all protein coding genes and are therefore enriched for regulatory elements (Larsen et al., 1992). Moreover, CGIs are sites of abnormal hypermethylation in neoplastic cells, which correlates with transcriptional repression of the associated gene promoters (Jones, 2002; Jones and Baylin, 2007). Consequently, CGIs represent a genomic fraction of considerable biological interest and one amenable to methylation analysis. Here we discuss the development and utility of novel reagents to characterise these sequences in primary human cells.

5.1 Generation and Characterisation of a Somatic CGI set

In this study we developed CAP, a novel chromatographic procedure for the selective enrichment of nonmethylated CGI sequences from bulk genomic DNA. This specific affinity is provided by a recombinant CXXC domain coupled at high density to a solid chromatographic matrix. This matrix provides selectivity for sequences bearing clusters of nonmethylated CpGs, making it ideally suited for the fractionation of CGIs. We prepared a somatic CGI set by CAP fractionation of human blood DNA. Extensive characterisation confirmed the enrichment of sequences with a base composition coherent with that of classical CGIs (Gardiner-Garden and Frommer, 1987; Takai and Jones, 2002). The complete set represents 17,387 discrete loci, which preferentially localise to gene rich regions of the genome. This set supersedes a previous library, which due to post-processing steps contains significantly fewer bona fide CGIs and a large proportion or repetitive elements (Cross et al., 1994; Heisler et al., 2005). Analysis of gene association suggests that the CXXC library contains approximately 25,000 CGIs of which approximately 60% have tenable sequence. PCR amplicons generated from these sequences were spotted onto microarray slides to allow further characterisation.

Consistent with previous reports, CGIs were found to preferentially localise to chromosome ends proximal to the telomeres (Craig and Bickmore, 1994). Despite a clear association with the transcription start sites of protein coding genes, approximately 51% were distal to annotated promoters. This finding was particularly interesting, as analysis of intra- or intergenic islands has previously been shown to identify unanticipated gene promoters (Gardiner-Garden and Frommer, 1987; Macleod et al., 1998). Moreover, it is possible that many of these islands localise to promoters of ncRNA genes. Accordingly, transcription of the ncRNAs *Xist* and *Air* initiate from CGI-promoters (Panning and Jaenisch, 1996; Sleutels et al., 2002; Wutz et al., 1997). A recent study identified a regulatory ncRNA which is transcribed from the *HOXC* locus (Rinn et al., 2007). All four *HOX* clusters contain an atypically high density of CGIs (Fig. 4.3-6d and data not shown). These results indicate a potential regulatory role for CGIs at these developmental gene loci.

One explanation for the distinct sequence composition of CGIs is transcriptional activity during embryogenesis (Antequera, 2003; Ponger et al., 2001). Accordingly, we found that approximately 78% of all CGIs associate with islands of H3K4me3 in human ES cells (Guenther et al., 2007). Moreover, non-promoter CGIs also display a high level of concordance (~64%) with these transcriptionally permissive regions of chromatin. This data supports the notion that CGIs persist due to a distinct chromatin state in embryonic cells. Moreover this affirms the suggestion that the majority of CGIs are sites of transcriptional initiation, at least during embryogenesis. A recent study determined that DNMT3L binding is inhibited by this H3K4me3 mark (Ooi et al., 2007). This proposes an intriguing scenario whereby transcriptionally active chromatin antagonizes the docking of the *de novo* methyltransferases. This would lead to the immunity of CGI sequence against *de novo* methylation during the phase of embryogenic reprogramming.

The CAP library was generated based on the clustering of nonmethylated CpG dinucleotides, rather than sequence composition alone. Comparison with a commonly used sequence based detection algorithm ($NCBI^{strict}$), confirmed a high level of concordance between the two methods. However, the CGI set contains approximately 23% of CGIs which were not identified by this algorithm. These sequences display a slightly reduced CpG density, suggesting that CAP provides a more biologically relevant method of CGI detection, without the imposition of arbitrary thresholds.

5.2 CGI Methylation in Human Cells

Somatic CGI methylation

Human CGI methylation is associated with important biological processes such as X-inactivation and parental imprinting. In these systems, methylated CGIs provide an epigenetic mark which is required for stable gene repression (Heard et al., 1997; Reik et al., 2001). There is also evidence to suggest that promoter-CGI methylation serves as a primary mechanism to silence transcription of germ-line specific genes (De Smet et al., 1999). In these cases, the functional significance of CGI methylation was realized following primary investigation of these systems. More recently however, DNA methylation profiling has been utilised as a means to identify hitherto undiscovered biological processes (Eckhardt et al., 2006; Oakes et al., 2007; Shen et al., 2007; Weber et al., 2005; Weber et al., 2007).

MeDIP analysis of ~16,000 human gene promoters determined that 3% of those associated with CGIs were heavily methylated in somatic cells (Weber et al., 2007). MAP array analysis suggested a somewhat higher proportion (8%) although this may have arisen due to the sub classification of CGI promoters as either 'intermediate' or 'high' in the former study (Weber et al., 2007). Strikingly, we observed a far higher frequency of methylation at non-promoter CGIs (~16%) with a three fold over representation at the 3' ends of genes relative to those associated with the TSS. Detailed analysis of 512 human CGIs identified ~9.2% which were heavily methylated in somatic cells, which is relatively consistent with the 11.6% identified in this study (1, 657 of 14,318; (Eckhardt et al., 2006)). Approximately, 3% (408 of 14,318) of CGIs were identified as concordantly methylated in all four tissues (408 of 14,318). It is likely however, that this represents an underestimate as complete methylation in blood DNA would exclude a proportion of CGIs from the library. Consistent with this idea, 22 methylated CGIs identified by Weber and coworkers were absent from CGI array (Weber et al., 2007).

Sequence Characteristics of Methylated Islands

An investigation of somatically methylated CGIs suggested that the most CpG-rich sequences were always nonmethylated (Eckhardt et al., 2006). Accordingly, analysis of gene promoter methylation indicated that 'intermediate' islands were more frequently methylated (21%) than those with a higher density of CpGs (3%; (Weber et al., 2007)). These findings suggest that elevated CpG-density may protect against DNA methylation at these sequences. However, analysis of the methylated component of the CGI set detected only a modest

reduction in CpG density (difference = 0.02 CpG[o/e]). Bock and coworkers identified a correlation between specific DNA sequence, simple repeats, DNA structure and CGI methylation status (Bock et al., 2006). Contrary to these findings, we found that repetitive elements were depleted in methylated islands, and did not show a significant difference in predicated base twist. We did observe a small but significant enrichment of (TGTG/CACA) simple repeats and base-stacking energy. Consequently, our data suggests that these sequence characteristics are not effective predictors of CGI methylation and the biological significance of these small effects is unclear. One issue that was not addressed in these studies was the chromosomal distribution of methylated CGIs. It is therefore interesting to suggest that modest compositional differences observed reflects a discrete genomic distribution. Consistent with this hypothesis, methylated islands were found to be disproportionately associated with telomeres which are known to have an atypical base composition (Riethman et al., 2004). Both CGI methylation and shortening of the telomeres have been associated with aging (Issa et al., 1996; Kwabi-Addo et al., 2007; Shawi and Autexier, 2008; Shiraishi et al., 1999). Since the majority of the DNA samples were prepared from individuals over sixty years of age, it is interesting to hypothesise that these chromosomal regions may be enriched for age correlated CGI methylation. Further investigation will be required to determine the validity of this hypothesis.

'Instructive' CGI Methylation

Neoplastic cells often display aberrant gene promoter methylation, and it has been proposed that these sites are enriched for H3K27me3 in human ES cells (Ohm et al., 2007; Schlesinger et al., 2007; Widschwendter et al., 2007). To address the possibility that 'normal' somatic CGI methylated shows an equivalent association, we compared somatically methylated CGIs with reported sites of H3K27me3 in human ES cells (Lee et al., 2006). We determined that CGIs associated with this modification were only slightly more likely to be methylated in somatic tissues (7.7 and 5.9% of methylated and all CGIs respectively). This suggests that 'normal' somatic CGI methylation is not preferentially targeted to sites of H3K27me3 modification and may therefore be mechanistically distinct from that described in cancer cells.

Tissue Specific CGI Methylation

Tissue specific differential CGI methylation has been documented in mouse and man (Eckhardt et al., 2006; Futscher et al., 2002; Grunau et al., 2000; Imamura et al., 2001; Shiota, 2004). However, the majority of these variably methylated islands, are

hypomethylated in the germ line and methylated in all cells of the soma (Brock et al., 1999; De Smet et al., 1999; Eckhardt et al., 2006; Kitamura et al., 2007; Oakes et al., 2007; Shen et al., 2007; Strichman-Almashanu et al., 2002; Weber et al., 2007). A relatively recent study characterizing CpG methylation profiles in human brain, identified distinct methylation patterns specific to individual brain regions (Ladd-Acosta et al., 2007). MAP array results from Blood, Brain, Muscle and Spleen identified between 5.7 and 8.3% of CGIs as methylated in different tissues. Further analysis confirmed that ~5% (711 of 14,318) of CGIs were differentially methylated within these somatic tissues. Interestingly, these islands are preferentially associated with developmental genes, including those of the *HOX* and *PAX* families (discussed in section 5.3).

5.3 Genes, Transcription and CGI Methylation

CGI Methylation and Transcriptional Regulation

There is extensive evidence to support a causative role for promoter-CGI methylation in transcriptional repression (see for example (De Smet et al., 1999; Stein et al., 1982; Weber et al., 2007)). Accordingly, we show that genes which escape inactivation on the Xi are atypically hypomethylated, upholding this dogma (Carrel and Willard, 2005; Weber et al., 2007). DNA methylation of the CpG rich promoters of human *MASPIN* and *GATA2* correlates with tissue specific gene silencing (Futscher et al., 2002; Song et al., 2005). In light of this evidence, it is tantalising to hypothesise that tissue specific CGI methylation may provide a mechanism to regulate tissue specific gene expression. However, the candidate gene *SEC31B*, was found to be repressed irrespective of the methylation status of its promoter-CGI. A comparison between differential CGI methylation and mined expression data also showed a poor correlation with transcriptional activity consistent with previous results (Oakes et al., 2007).

A potential difficulty in interpreting the role of CGI methylation in transcriptional repression is that methylation profiles generated in this study could not be compared with expression data from the same biological samples. Secondly, many of the differentially methylated islands are located within intragenic regions and the consequence of methylation at these sites is unclear (Grunau et al., 2000; Kitamura et al., 2007; Ladd-Acosta et al., 2007). Finally, expression may be restricted to a subset of cells within a mixed population comprising the analysed tissue. To determine the validity and contribution of these factors a more extensive expression analysis will be required. RNA and DNA prepared from the same

source will provide a more robust comparison. However, this was not possible at the time of investigation due to the limited availability of high quality human tissue.

CGI Methylation: Initiation or Maintenance of Transcriptional Repression?

Evidence from X inactivation and transgenic studies indicates that transcriptional repression precedes promoter-CGI methylation (see Introduction). Analysis of *SEC31B* and mined expression data suggests that differential methylation frequently occurs at constitutively silenced genes. This indicates that CGI methylation may follow, rather than initiate, transcriptional repression. Two models could account for this possibility. 1) The absence of TFs at silenced promoters could facilitate transient *de novo* methylation due to the lack of steric hindrance. This possibility would align with the notion that methylation is the basal state of the genome which is excluded from specific regions by the presence of bound factors. 2) DNMT recruitment could be mediated by initial repressive events, to irrevocably silence transcription of the associated gene. Combined analysis of RNA polymerase occupancy, histone modifications, expression and acquisition of CGI methylation during development may help to resolve this issue.

Differential CGI Methylation and Gene Function

Developmental gene functions such as neurogenesis, segmentatation specification and mesoderm development were prominent amongst genes associated with differentially methylated CGIs. Specifically, association of differential methylation was approximately three fold overrepresented at the Homeobox developmental regulatory genes. Moreover, out of 79 CGIs associated with the four HOX gene clusters, approximately 25% (22 of 79) were methylated in at least on of the four tissues tested. These highly conserved genes dictate the positional identities of cells within the human body and therefore represent key regulators of mammalian development. It is interesting to note, that expanding the panel of tissues will likely uncover additional methylated islands within these important developmental gene loci. Further characterisation of these important regulatory genes will hopefully provide further insight into the function of CGI methylation, during cellular differentiation.

CGI methylation: A note regarding PAX6

PAX6 is a transcription factor required for ocular and neural development and it's expression is temporally and spatially partitioned within the mammalian brain (Kleinjan et al., 2004). The observation that an intragenic CGI sequence was specifically methylated in brain suggests an interesting hypothesis (Fig. 4.3-6b). CGI methylation may serve to restrict gene

expression to a specific cell type within a tissue. In the case of *PAX6*, this mechanism may be required to prevent ectopic expression in cells containing all the TFs required for efficient expression. The effect of perturbed *PAX6* expression is illustrated in the congenital disorder Aniridia, where lack of *PAX6* leads to improper eye formation (Kodama and Eguchi, 1994; van Raamsdonk and Tilghman, 2000).

5.4 Composite methylation

Bisulfite sequence analysis determined that several MAP enriched CGIs display a composite methylation pattern. Several explanations could account for this mixed population of heavily methylated and nonmethylated DNA molecules. At the highest level, the composite pattern may indicate individual variation in the pooled DNA. To address this possibility, methylation analysis was carried out on muscle DNA from three individuals. Accordingly, individual C displayed a considerably higher level of methylation than individuals A and B (Fig. 4.3-5b). Therefore, such polymorphic methylation may be acquired during the course of human aging, although this remains a contentious issue. Whilst several studies have identified an increased incidence of CGI methylation with age, others have failed to confirm these results (Eckhardt et al., 2006; Kwabi-Addo et al., 2007; Ladd-Acosta et al., 2007). Alternatively, polymorphic methylation may occur as a random stochastic event or as a consequence of some underlying disease state. Individual polymorphic methylation is of significant interest as it may represent an epigenetic determinant of phenotypic variability. A large scale screen of individual CGI methylation will provide insight into the extent of variation and the relative contribution of each of these factors.

A second possibility is that methylation patterns are distinct within tissues resulting from differential methylation of constituent cell types. As previously discussed, we identified 5% of CGIs which display variable methylation patterns between four somatic tissues. A recent study addressed the methylation status of 1,505 CpG sites in a panel of individual matched brain regions (Cerebrum, Cerebellum and Pons). Interestingly, region specific methylation patterns were consistently observed and the authors propose that this is indicative of regional specialization (Ladd-Acosta et al., 2007). It is therefore conceivable that variable methylation patterns exist between functionally distinct cell types. Here we show that the promoter-CGI of *SEC31B* was more heavily methylated in mononuclear cells than granulocyte cells prepared from whole blood. It is interesting to propose that cell specific methylation patterns may serve to determine or maintain expression profiles following tissue differentiation.

Monoallelic CGI methylation is frequently associated with imprinted gene loci and an equivalent phenomenon could account for composite CGI methylation (Reik, 2007). Accordingly, analysis of predicted CGIs on chromosome arm 21q identified ~5% (7 of 149) of CGIs which were compositely methylated in peripheral blood DNA. Three of these islands were monoallelically methylated as determined by SNP analysis (Yamada et al., 2004). In the absence of parent specific SNP data we were unable to test this hypothesis directly. Although, comparison of genes identified as being monoallelically expressed in blood cells, showed a poor association with the methylated CGIs identified here (Gimelbrant et al., 2007). Consequently, we conclude that monoallelic methylation is not a major contributor to this phenomenon; however more extensive analysis is required.

5.5 Future Work

Second Generation CGI Libraries

CXXC affinity purification is highly effective at selecting DNA fragments containing clusters of non-methylated CpG sites. This was illustrated by the characterisation of clone inserts from the CGI library in Chapter 3. Despite the coverage of the CGI library however, there are several aspects which could have been improved. 1) The use of conventional sequencing chemistry led to a reduced coverage of the library due to bias against G+C rich sequence composition. 2) MseI restriction provides an effective step in the preparation of a CGI fraction as has been discussed previously. However, its application in the preparation of the CXXC fraction resulted in fragmentation of many of the CGIs. Moreover many islands were only represented by flanking MseI fragments which were useful for CGI localisation but rather ineffective as arrayed probes. 3). Due to practical considerations, the library was generated from blood, rather than sperm. This allowed the selection of hypomethylated CGIs whilst avoiding the purification of certain repeat sequences which would provide efficient ligands for the CXXC matrix in sperm. Consequently, the library is depleted for all CGIs which are heavily methylated in blood, as confirmed by the absence of previously identified methylated CGI sequences (Futscher et al., 2002; Weber et al., 2007).

These factors result from the limitations of conventional sequencing, either by the inability to sequence specific base compositions or to generate sufficient reads to provide coverage of complex DNA fractions. The use of Solexa sequencing circumvents these problems by producing tens of millions of compositionally unbiased sequences. To this end, CXXC

fractions have been generated from sonicated blood and sperm DNA, and sequenced using the solexa technology (Wellcome Trust Sanger institute's sequencing team, under the direction of Dr Julian Parkhill). The CAP fractions are currently being investigated; however the preliminary sequences are coherent with the composition of classical CGIs (Fig. 5.5-1a-c). There are significant differences between the blood and sperm preparations, such as the promoter associated CGI of *CATSPER2* (Cation Channel: Sperm Associated 2; Fig. 5.5-1c). *CATSPER2* belongs to a gene family which is expressed specifically in sperm and therefore resembles genes of the *MAGE* family (De Smet et al., 1999; Li et al., 2007b). Replication and refinement of this data will provide a complete unbiased set of CGIs. Moreover, analysis of CAP preparations from blood and sperm DNA will hopefully provide an insight into the acquisition of CGI methylation in somatic cell lineages.

Functional Role of Tissue Specific CGI Methylation

Several studies including this one have identified tissue specific methylation patterns (Eckhardt et al., 2006; Futscher et al., 2002; Imamura et al., 2001; Ladd-Acosta et al., 2007; Shiota et al., 2002). However, with the exception of *MASPIN*, there is little direct evidence to suggest that tissue specific expression patterns are dictated by differential methylation events (Futscher et al., 2002). To address this problem it would be informative to relate methylation patterns to transcription status derived from the same biological samples. An alternative elegant technical approach is provided by sequential ChIP-bisulfite sequencing (D'Alessio et al., 2007). This allows the determination of DNA methylation patterns in chromatin associated with active or inactive marks or specific transcription factors. The method is particularly versatile as it allows direct comparison of DNA methylation and transcriptional permissivity at single molecule resolution. Cell and allele[xxxv] specific methylation can be resolved by this approach.

How Variable are CGI Methylation Patterns?

As discussed, there are several possible explanations to account for the composite methylation patterns identified by bisulfite sequence analysis in this study. To dissect the relative contribution of individual variation (age, normal variability, disease, stochastic and sex specific) and tissue specific effects on the methylation patterns, a more extensive analysis is required. A MAP (or MeDIP) microarray screen of human tissues from a large number of individuals will be required to account for the relative contribution of each of

[xxxv] Characterisation of allelic methylation requires the association of informative SNPs and, in the case of parental imprinting parent specific SNPs.

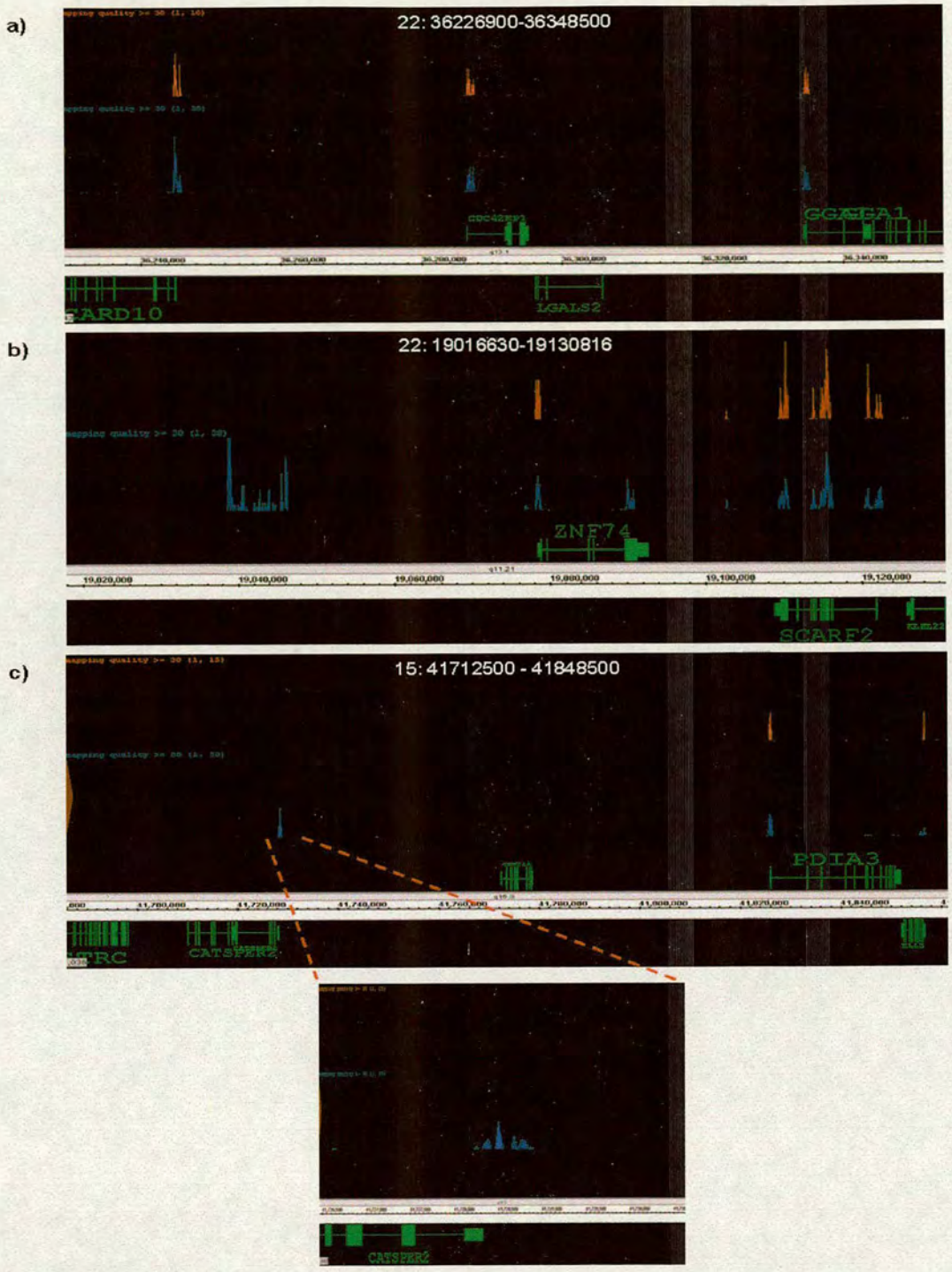


Figure 5.1-1. Mapped Solexa Sequence from CAP purified Blood and Sperm DNA
(**a-c**) Solexa sequence from CAP fractionated Blood (Red) and Sperm (Blue) DNA, mapped to three regions of the human genome (NCBI build 36). Each region depicts nonmethylated CGIs purified from blood and sperm and several specific to sperm alone (panels **b** and **c**). (**c**) A region of chromosome 15 and an enlarged view of the sperm specific CGI associated with the promoter of *CATSPER2*. Chromosomal region is indicated in the top centre of each panel following the convention of chromosome number: start-stop. Genes (green bars) are transcribed left-right on the sense strand (upper) and right-left on the antisense strand (lower).

these components. This will provide an effective means to fully appreciate the source of these methylation patterns and to identify the regulatory functions that they may provide.

5.6 Concluding Remarks

Selection of the majority of human CGIs sequences based on their intrinsic molecular properties has allowed for a comprehensive characterisation of this interesting genomic fraction. This has provided a refined understanding of the distribution and abundance of CGIs within the context of the genome and individual genes. Moreover, it serve as an analytical tool which has provided tantalising insights into the function of CGI methylation and development. Future analysis based on the generalizations drawn here, will hopefully provide insight into the biological function of these sequences in cell identity and gene regulation.

References

Agger, K., Cloos, P.A., Christensen, J., Pasini, D., Rose, S., Rappsilber, J., Issaeva, I., Canaani, E., Salcini, A.E., and Helin, K. (2007). UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. Nature *449*, 731-734.

Ahmad, K., and Henikoff, S. (2002). The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. Mol Cell *9*, 1191-1200.

Allen, M.D., Grummitt, C.G., Hilcenko, C., Min, S.Y., Tonkin, L.M., Johnson, C.M., Freund, S.M., Bycroft, M., and Warren, A.J. (2006). Solution structure of the nonmethyl-CpG-binding CXXC domain of the leukaemia-associated MLL histone methyltransferase. Embo J *25*, 4503-4512.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U., and Zoghbi, H.Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet *23*, 185-188.

Ansari, K.I., Mishra, B.P., and Mandal, S.S. (2008). Human CpG binding protein interacts with MLL1, MLL2 and hSet1 and regulates Hox gene expression. Biochim Biophys Acta *1779*, 66-73.

Antequera, F. (2003). Structure, function and evolution of CpG island promoters. Cell Mol Life Sci *60*, 1647-1658.

Antequera, F., and Bird, A. (1993). Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A *90*, 11995-11999.

Antequera, F., and Bird, A. (1999). CpG islands as genomic footprints of promoters that are associated with replication origins. Curr Biol *9*, R661-667.

Antequera, F., Boyes, J., and Bird, A. (1990). High levels of de novo methylation and altered chromatin structure at CpG islands in cell lines. Cell *62*, 503-514.

Antequera, F., Macleod, D., and Bird, A.P. (1989). Specific protection of methylated CpGs in mammalian nuclei. Cell *58*, 509-517.

Aravin, A.A., and Bourc'his, D. (2008). Small RNA guides for de novo DNA methylation in mammalian germ cells. Genes Dev *22*, 970-975.

Ayton, P.M., Chen, E.H., and Cleary, M.L. (2004). Binding to nonmethylated CpG DNA is essential for target recognition, transactivation, and myeloid transformation by an MLL oncoprotein. Mol Cell Biol *24*, 10470-10478.

Bader, S., Walker, M., and Harrison, D. (2000). Most microsatellite unstable sporadic colorectal carcinomas carry MBD4 mutations. British journal of cancer *83*, 1646-1649.

Bannister, A.J., Zegerman, P., Partridge, J.F., Miska, E.A., Thomas, J.O., Allshire, R.C., and Kouzarides, T. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature *410*, 120-124.

Barr, H., Hermann, A., Berger, J., Tsai, H.H., Adie, K., Prokhortchouk, A., Hendrich, B., and Bird, A. (2007). Mbd2 contributes to DNA methylation-directed repression of the Xist gene. Mol Cell Biol *27*, 3750-3757.

Barry, C., Faugeron, G., and Rossignol, J.L. (1993). Methylation induced premeiotically in Ascobolus: coextension with DNA repeat lengths and effect on transcript elongation. Proc Natl Acad Sci U S A *90*, 4557-4561.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Becker, P.B., Ruppert, S., and Schutz, G. (1987). Genomic footprinting reveals cell type-specific DNA binding of ubiquitous factors. Cell *51*, 435-443.

Bell, A.C., and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature *405*, 482-485.

Bender, J. (2004). DNA methylation and epigenetics. Annual review of plant biology *55*, 41-68.

Benetti, R., Gonzalo, S., Jaco, I., Munoz, P., Gonzalez, S., Schoeftner, S., Murchison, E., Andl, T., Chen, T., Klatt, P., *et al.* (2008). A mammalian microRNA cluster controls DNA methylation and telomere recombination via Rbl2-dependent regulation of DNA methyltransferases. Nat Struct Mol Biol *15*, 268-279.

Berger, J., Sansom, O., Clarke, A., and Bird, A. (2007). MBD2 is required for correct spatial gene expression in the gut. Mol Cell Biol *27*, 4049-4057.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R., *et al.* (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell *120*, 169-181.

Bernstein, B.E., Meissner, A., and Lander, E.S. (2007). The mammalian epigenome. Cell *128*, 669-681.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., *et al.* (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., *et al.* (2004). Global identification of human transcribed sequences with genome tiling arrays. Science *306*, 2242-2246.

Bestor, T.H. (1992). Activation of mammalian DNA methyltransferase by cleavage of a Zn binding regulatory domain. Embo J *11*, 2611-2617.

Bestor, T.H., and Tycko, B. (1996). Creation of genomic methylation patterns. Nat Genet *12*, 363-367.

Bhattacharya, S.K., Ramchandani, S., Cervoni, N., and Szyf, M. (1999). A mammalian protein with specific demethylase activity for mCpG DNA. Nature *397*, 579-583.

Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E.W., Wu, B., Doucet, D., Thomas, N.J., Wang, Y., Vollmer, E., *et al.* (2006). High-throughput DNA methylation profiling using universal bead arrays. Genome Res *16*, 383-393.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev *16*, 6-21.

Bird, A. (2007). Perceptions of epigenetics. Nature *447*, 396-398.

Bird, A., Taggart, M., Frommer, M., Miller, O.J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell *40*, 91-99.

Bird, A., Tate, P., Nan, X., Campoy, J., Meehan, R., Cross, S., Tweedie, S., Charlton, J., and Macleod, D. (1995). Studies of DNA methylation in animals. J Cell Sci Suppl *19*, 37-39.

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res *8*, 1499-1504.

Bird, A.P. (1987). CpG Islands as gene markers in the Vertebrate Nucleus. TIG *3*, 342-347.

Bird, A.P. (1995). Gene number, noise reduction and biological complexity. Trends Genet *11*, 94-100.

Bird, A.P., and Southern, E.M. (1978). Use of restriction enzymes to study eukaryotic DNA methylation: I. The methylation pattern in ribosomal DNA from Xenopus laevis. J Mol Biol *118*, 27-47.

Bird, A.P., and Taggart, M.H. (1980). Variable patterns of total DNA and rDNA methylation in animals. Nucleic Acids Res *8*, 1485-1497.

Bird, A.P., and Wolffe, A.P. (1999). Methylation-induced repression--belts, braces, and chromatin. Cell *99*, 451-454.

Birke, M., Schreiner, S., Garcia-Cuellar, M.P., Mahr, K., Titgemeyer, F., and Slany, R.K. (2002). The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and discriminates against methylation. Nucleic Acids Res *30*, 958-965.

Birney, E., Andrews, T.D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., *et al.* (2004). An overview of Ensembl. Genome Res *14*, 925-928.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., *et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816.

Bock, C., Paulsen, M., Tierling, S., Mikeska, T., Lengauer, T., and Walter, J. (2006). CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. PLoS Genet *2*, e26.

Bock, C., Walter, J., Paulsen, M., and Lengauer, T. (2007). CpG island mapping by epigenome prediction. PLoS computational biology *3*, e110.

Boumil, R.M., Ogawa, Y., Sun, B.K., Huynh, K.D., and Lee, J.T. (2006). Differential methylation of Xite and CTCF sites in Tsix mirrors the pattern of X-inactivation choice in mice. Mol Cell Biol *26*, 2109-2117.

Bourc'his, D., and Bestor, T.H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature *431*, 96-99.

Bourc'his, D., Xu, G.L., Lin, C.S., Bollman, B., and Bestor, T.H. (2001). Dnmt3L and the establishment of maternal genomic imprints. Science *294*, 2536-2539.

Bovee, D., Zhou, Y., Haugen, E., Wu, Z., Hayden, H.S., Gillett, W., Tuzun, E., Cooper, G.M., Sampas, N., Phelps, K., *et al.* (2008). Closing gaps in the human genome with fosmid resources generated from multiple individuals. Nat Genet *40*, 96-101.

Bowen, N.J., Fujita, N., Kajita, M., and Wade, P.A. (2004). Mi-2/NuRD: multiple complexes for many purposes. Biochim Biophys Acta *1677*, 52-57.

Boyes, J., and Bird, A. (1991). DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. Cell *64*, 1123-1134.

Boyes, J., and Bird, A. (1992). Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. Embo J *11*, 327-333.

Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A., and Cedar, H. (1994). Sp1 elements protect a CpG island from de novo methylation. Nature *371*, 435-438.

Brenner, C., Deplus, R., Didelot, C., Loriot, A., Vire, E., De Smet, C., Gutierrez, A., Danovi, D., Bernard, D., Boon, T., *et al.* (2005). Myc represses transcription through recruitment of DNA methyltransferase corepressor. Embo J *24*, 336-346.

Brock, G.J., and Bird, A. (1997). Mosaic methylation of the repeat unit of the human ribosomal RNA genes. Hum Mol Genet *6*, 451-456.

Brock, G.J., Charlton, J., and Bird, A. (1999). Densely methylated sequences that are preferentially localized at telomere-proximal regions of human chromosomes. Gene *240*, 269-277.

Brock, G.J., Huang, T.H., Chen, C.M., and Johnson, K.J. (2001). A novel technique for the identification of CpG islands exhibiting altered methylation patterns (ICEAMP). Nucleic Acids Res *29*, E123.

Brockdorff, N. (2002). X-chromosome inactivation: closing in on proteins that bind Xist RNA. Trends Genet *18*, 352-358.

Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G., Grompe, M., Tonlorenzi, R., and Willard, H.F. (1991). A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature *349*, 38-44.

Brown, C.J., and Greally, J.M. (2003). A stain upon the silence: genes escaping X inactivation. Trends Genet *19*, 432-438.

Bruniquel, D., and Schwartz, R.H. (2003). Selective, stable demethylation of the interleukin-2 gene enhances transcription by an active process. Nature immunology *4*, 235-240.

Campanero, M.R., Armstrong, M.I., and Flemington, E.K. (2000). CpG methylation as a mechanism for the regulation of E2F activity. Proc Natl Acad Sci U S A *97*, 6481-6486.

Cao, R., and Zhang, Y. (2004). The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3. Curr Opin Genet Dev *14*, 155-164.

Carlone, D.L., Hart, S.R., Ladd, P.D., and Skalnik, D.G. (2002). Cloning and characterization of the gene encoding the mouse homologue of CpG binding protein. Gene *295*, 71-77.

Carrel, L., Clemson, C.M., Dunn, J.M., Miller, A.P., Hunt, P.A., Lawrence, J.B., and Willard, H.F. (1996). X inactivation analysis and DNA methylation studies of the ubiquitin activating enzyme E1 and PCTAIRE-1 genes in human and mouse. Hum Mol Genet *5*, 391-401.

Carrel, L., and Willard, H.F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature *434*, 400-404.

Carrozza, M.J., Li, B., Florens, L., Suganuma, T., Swanson, S.K., Lee, K.K., Shia, W.J., Anderson, S., Yates, J., Washburn, M.P., *et al.* (2005). Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell *123*, 581-592.

Cavalli, G., and Paro, R. (1998). The Drosophila Fab-7 chromosomal element conveys epigenetic inheritance during mitosis and meiosis. Cell *93*, 505-518.

Chaillet, J.R., Vogt, T.F., Beier, D.R., and Leder, P. (1991). Parental-specific methylation of an imprinted transgene is established during gametogenesis and progressively changes during embryogenesis. Cell *66*, 77-83.

Chamberlain, S.J., Yee, D., and Magnuson, T. (2008). Polycomb repressive complex 2 is dispensable for maintenance of embryonic stem cell pluripotency. Stem cells (Dayton, Ohio) *26*, 1496-1505.

Chan, S.W., Henderson, I.R., and Jacobsen, S.E. (2005). Gardening the genome: DNA methylation in Arabidopsis thaliana. Nature reviews *6*, 351-360.

Chen, C.M., Chen, H.L., Hsiau, T.H., Hsiau, A.H., Shi, H., Brock, G.J., Wei, S.H., Caldwell, C.W., Yan, P.S., and Huang, T.H. (2003a). Methylation target array for rapid analysis of CpG island hypermethylation in multiple tissue genomes. Am J Pathol *163*, 37-45.

Chen, T., Hevi, S., Gay, F., Tsujimoto, N., He, T., Zhang, B., Ueda, Y., and Li, E. (2007). Complete inactivation of DNMT1 leads to mitotic catastrophe in human cancer cells. Nat Genet *39*, 391-396.

Chen, T., Ueda, Y., Dodge, J.E., Wang, Z., and Li, E. (2003b). Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. Mol Cell Biol *23*, 5594-5605.

Chesnokov, I.N., and Schmid, C.W. (1995). Specific Alu binding protein from human sperm chromatin prevents DNA methylation. J Biol Chem *270*, 18539-18542.

Chuang, L.S., Ian, H.I., Koh, T.W., Ng, H.H., Xu, G., and Li, B.F. (1997). Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. Science *277*, 1996-2000.

Clemson, C.M., McNeil, J.A., Willard, H.F., and Lawrence, J.B. (1996). XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. The Journal of cell biology *132*, 259-275.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature *452*, 215-219.

Colot, V., and Rossignol, J.L. (1999). Eukaryotic DNA methylation as an evolutionary device. Bioessays *21*, 402-411.

Cooper, D.N., and Krawczak, M. (1990). The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. Human genetics *85*, 55-74.

Cooper, D.N., Taggart, M.H., and Bird, A.P. (1983). Unmethylated domains in vertebrate DNA. Nucleic Acids Res *11*, 647-658.

Craig, J.M., and Bickmore, W.A. (1994). The distribution of CpG islands in mammalian chromosomes. Nat Genet *7*, 376-382.

Cross, S.H., Charlton, J.A., Nan, X., and Bird, A.P. (1994). Purification of CpG islands using a methylated DNA binding column. Nat Genet *6*, 236-244.

Cross, S.H., Clark, V.H., and Bird, A.P. (1999). Isolation of CpG islands from large genomic clones. Nucleic Acids Res *27*, 2099-2107.

Cross, S.H., Clark, V.H., Simmen, M.W., Bickmore, W.A., Maroon, H., Langford, C.F., Carter, N.P., and Bird, A.P. (2000). CpG island libraries from human chromosomes 18 and 22: landmarks for novel genes. Mamm Genome *11*, 373-383.

Cross, S.H., Lee, M., Clark, V.H., Craig, J.M., Bird, A.P., and Bickmore, W.A. (1997a). The chromosomal distribution of CpG islands in the mouse: evidence for genome scrambling in the rodent lineage. Genomics *40*, 454-461.

Cross, S.H., Meehan, R.R., Nan, X., and Bird, A. (1997b). A component of the transcriptional repressor MeCP1 shares a motif with DNA methyltransferase and HRX proteins. Nat Genet *16*, 256-259.

Csankovszki, G., Nagy, A., and Jaenisch, R. (2001). Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. The Journal of cell biology *153*, 773-784.

Csankovszki, G., Panning, B., Bates, B., Pehrson, J.R., and Jaenisch, R. (1999). Conditional deletion of Xist disrupts histone macroH2A localization but not maintenance of X inactivation. Nat Genet *22*, 323-324.

Cuadrado, M., Sacristan, M., and Antequera, F. (2001). Species-specific organization of CpG island promoters at mammalian homologous genes. EMBO reports *2*, 586-592.

D'Alessio, A.C., Weaver, I.C., and Szyf, M. (2007). Acetylation-induced transcription is required for active DNA demethylation in methylation-silenced genes. Mol Cell Biol *27*, 7462-7474.

Daniel, J.M., Spring, C.M., Crawford, H.C., Reynolds, A.B., and Baig, A. (2002). The p120(ctn)-binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides. Nucleic Acids Res *30*, 2911-2919.

Daniels, R., Lowell, S., Bolton, V., and Monk, M. (1997). Transcription of tissue-specific genes in human preimplantation embryos. Human reproduction (Oxford, England) *12*, 2251-2256.

Davis, T.L., Trasler, J.M., Moss, S.B., Yang, G.J., and Bartolomei, M.S. (1999). Acquisition of the H19 methylation imprint occurs differentially on the parental alleles during spermatogenesis. Genomics *58*, 18-28.

Davis, T.L., Yang, G.J., McCarrey, J.R., and Bartolomei, M.S. (2000). The H19 methylation imprint is erased and re-established differentially on the parental alleles during male germ cell development. Hum Mol Genet *9*, 2885-2894.

de Napoles, M., Mermoud, J.E., Wakao, R., Tang, Y.A., Endoh, M., Appanah, R., Nesterova, T.B., Silva, J., Otte, A.P., Vidal, M., *et al.* (2004). Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. Developmental cell *7*, 663-676.

De Smet, C., Lurquin, C., Lethe, B., Martelange, V., and Boon, T. (1999). DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter. Mol Cell Biol *19*, 7327-7335.

Delgado, S., Gomez, M., Bird, A., and Antequera, F. (1998). Initiation of DNA replication at CpG islands in mammalian chromosomes. Embo J *17*, 2426-2435.

Di Croce, L., Raker, V.A., Corsaro, M., Fazi, F., Fanelli, M., Faretta, M., Fuks, F., Lo Coco, F., Kouzarides, T., Nervi, C., *et al.* (2002). Methyltransferase recruitment and DNA hypermethylation of target promoters by an oncogenic transcription factor. Science *295*, 1079-1082.

Dodge, J.E., Okano, M., Dick, F., Tsujimoto, N., Chen, T., Wang, S., Ueda, Y., Dyson, N., and Li, E. (2005). Inactivation of Dnmt3b in mouse embryonic fibroblasts results in DNA hypomethylation, chromosomal instability, and spontaneous immortalization. J Biol Chem *280*, 17986-17991.

Dou, Y., Milne, T.A., Tackett, A.J., Smith, E.R., Fukuda, A., Wysocka, J., Allis, C.D., Chait, B.T., Hess, J.L., and Roeder, R.G. (2005). Physical association and coordinate function of the H3 K4 methyltransferase MLL1 and the H4 K16 acetyltransferase MOF. Cell *121*, 873-885.

Dunican, D.S., Ruzov, A., Hackett, J.A., and Meehan, R.R. (2008). xDnmt1 regulates transcriptional silencing in pre-MBT Xenopus embryos independently of its catalytic function. Development *135*, 1295-1302.

Eads, C.A., Danenberg, K.D., Kawakami, K., Saltz, L.B., Blake, C., Shibata, D., Danenberg, P.V., and Laird, P.W. (2000). MethyLight: a high-throughput assay to measure DNA methylation. Nucleic Acids Res *28*, E32.

Eckhardt, F., Beck, S., Gut, I.G., and Berlin, K. (2004). Future potential of the Human Epigenome Project. Expert review of molecular diagnostics *4*, 609-618.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A., *et al.* (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet *38*, 1378-1385.

Eden, A., Gaudet, F., Waghmare, A., and Jaenisch, R. (2003). Chromosomal instability and tumors promoted by DNA hypomethylation. Science *300*, 455.

Ehrlich, M. (2003). The ICF syndrome, a DNA methyltransferase 3B deficiency and immunodeficiency disease. Clinical immunology (Orlando, Fla *109*, 17-28.

Ehrlich, M., Gama-Sosa, M.A., Huang, L.H., Midgett, R.M., Kuo, K.C., McCune, R.A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic Acids Res *10*, 2709-2721.

ENCODE Project Consortium, B.E., Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermüller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi

U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaöz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Löytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JN, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrímsdóttir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636-640.

Epstein, N.D., Karlsson, S., O'Brien, S., Modi, W., Moulton, A., and Nienhuis, A.W. (1987). A new moderately repetitive DNA sequence family of novel organization. Nucleic Acids Res *15*, 2327-2341.

Estecio, M.R., Yan, P.S., Ibrahim, A.E., Tellez, C.S., Shen, L., Huang, T.H., and Issa, J.P. (2007). High-throughput methylation profiling by MCA coupled to CpG island microarray. Genome Res *17*, 1529-1536.

Ewing, B., and Green, P. (2000). Analysis of expressed sequence tags indicates 35,000 human genes. Nat Genet *25*, 232-234.

Fan, J.B., Gunderson, K.L., Bibikova, M., Yeakley, J.M., Chen, J., Wickham Garcia, E., Lebruska, L.L., Laurent, M., Shen, R., and Barker, D. (2006). Illumina universal bead arrays. Methods in enzymology *410*, 57-73.

Filippova, G.N. (2008). Genetics and epigenetics of the multifunctional protein CTCF. Current topics in developmental biology *80*, 337-360.

Fouse, S.D., Shen, Y., Pellegrini, M., Cole, S., Meissner, A., Van Neste, L., Jaenisch, R., and Fan, G. (2008). Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. Cell stem cell *2*, 160-169.

Fraga, M.F., Ballestar, E., Montoya, G., Taysavang, P., Wade, P.A., and Esteller, M. (2003). The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. Nucleic Acids Res *31*, 1765-1774.

Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suner, D., Cigudosa, J.C., Urioste, M., Benitez, J., *et al.* (2005). Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A *102*, 10604-10609.

Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A *89*, 1827-1831.

Fujita, N., Shimotake, N., Ohki, I., Chiba, T., Saya, H., Shirakawa, M., and Nakao, M. (2000). Mechanism of transcriptional regulation by methyl-CpG binding protein MBD1. Mol Cell Biol *20*, 5107-5118.

Fujita, N., Takebayashi, S., Okumura, K., Kudo, S., Chiba, T., Saya, H., and Nakao, M. (1999). Methylation-mediated transcriptional silencing in euchromatin by methyl-CpG binding protein MBD1 isoforms. Mol Cell Biol *19*, 6415-6426.

Fukasawa, K. (2005). Centrosome amplification, chromosome instability and cancer development. Cancer letters *230*, 6-19.

Fuks, F., Burgers, W.A., Brehm, A., Hughes-Davies, L., and Kouzarides, T. (2000). DNA methyltransferase Dnmt1 associates with histone deacetylase activity. Nat Genet *24*, 88-91.

Fuks, F., Hurd, P.J., Deplus, R., and Kouzarides, T. (2003a). The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. Nucleic Acids Res *31*, 2305-2312.

Fuks, F., Hurd, P.J., Wolf, D., Nan, X., Bird, A.P., and Kouzarides, T. (2003b). The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. J Biol Chem *278*, 4035-4040.

Futscher, B.W., Oshiro, M.M., Wozniak, R.J., Holtan, N., Hanigan, C.L., Duan, H., and Domann, F.E. (2002). Role for DNA methylation in the control of cell type specific maspin expression. Nat Genet *31*, 175-179.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J Mol Biol *196*, 261-282.

Gardiner-Garden, M., and Frommer, M. (1994). Transcripts and CpG islands associated with the pro-opiomelanocortin gene and other neurally expressed genes. Journal of molecular endocrinology *12*, 365-382.

Gaston, K., and Fried, M. (1995). CpG methylation has differential effects on the binding of YY1 and ETS proteins to the bi-directional promoter of the Surf-1 and Surf-2 genes. Nucleic Acids Res *23*, 901-909.

Gautsch, J.W., and Wilson, M.C. (1983). Delayed de novo methylation in teratocarcinoma suggests additional tissue-specific mechanisms for controlling gene expression. Nature *301*, 32-37.

Ge, Y.Z., Pu, M.T., Gowher, H., Wu, H.P., Ding, J.P., Jeltsch, A., and Xu, G.L. (2004). Chromatin targeting of de novo DNA methyltransferases by the PWWP domain. J Biol Chem *279*, 25447-25454.

Gebhard, C., Schwarzfischer, L., Pham, T.H., Andreesen, R., Mackensen, A., and Rehli, M. (2006a). Rapid and sensitive detection of CpG-methylation using methyl-binding (MB)-PCR. Nucleic Acids Res *34*, e82.

Gebhard, C., Schwarzfischer, L., Pham, T.H., Schilling, E., Klug, M., Andreesen, R., and Rehli, M. (2006b). Genome-wide profiling of CpG methylation identifies novel targets of aberrant hypermethylation in myeloid leukemia. Cancer Res *66*, 6118-6128.

Georgel, P.T., Horowitz-Scherer, R.A., Adkins, N., Woodcock, C.L., Wade, P.A., and Hansen, J.C. (2003). Chromatin compaction by human MeCP2. Assembly of novel secondary chromatin structures in the absence of DNA methylation. J Biol Chem *278*, 32181-32188.

Gilbert, S.L., and Sharp, P.A. (1999). Promoter-specific hypoacetylation of X-inactivated genes. Proc Natl Acad Sci U S A *96*, 13825-13830.

Gimelbrant, A., Hutchinson, J.N., Thompson, B.R., and Chess, A. (2007). Widespread monoallelic expression on human autosomes. Science *318*, 1136-1140.

Gitan, R.S., Shi, H., Chen, C.M., Yan, P.S., and Huang, T.H. (2002). Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis. Genome Res *12*, 158-164.

Goll, M.G., Kirpekar, F., Maggert, K.A., Yoder, J.A., Hsieh, C.L., Zhang, X., Golic, K.G., Jacobsen, S.E., and Bestor, T.H. (2006). Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. Science *311*, 395-398.

Gonzalgo, M.L., and Jones, P.A. (1997). Rapid quantitation of methylation differences at specific sites using methylation-sensitive single nucleotide primer extension (Ms-SNuPE). Nucleic Acids Res *25*, 2529-2531.

Gowher, H., and Jeltsch, A. (2001). Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive manner and also methylates non-CpG [correction of non-CpA] sites. J Mol Biol *309*, 1201-1208.

Gowher, H., Liebert, K., Hermann, A., Xu, G., and Jeltsch, A. (2005). Mechanism of stimulation of catalytic activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. J Biol Chem *280*, 13341-13348.

Graves, J.A. (2006). Sex chromosome specialization and degeneration in mammals. Cell *124*, 901-914.

Grewal, S.I., and Elgin, S.C. (2007). Transcription and RNA interference in the formation of heterochromatin. Nature *447*, 399-406.

Grunau, C., Hindermann, W., and Rosenthal, A. (2000). Large-scale methylation analysis of human genomic DNA reveals tissue-specific differences between the methylation profiles of genes and pseudogenes. Hum Mol Genet *9*, 2651-2663.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Gutierrez, A., and Sommer, R.J. (2004). Evolution of dnmt-2 and mbd-2-like genes in the free-living nematodes Pristionchus pacificus, Caenorhabditis elegans and Caenorhabditis briggsae. Nucleic Acids Res *32*, 6388-6396.

Guy, J., Gan, J., Selfridge, J., Cobb, S., and Bird, A. (2007). Reversal of neurological defects in a mouse model of Rett syndrome. Science *315*, 1143-1147.

Guy, J., Hendrich, B., Holmes, M., Martin, J.E., and Bird, A. (2001). A mouse Mecp2-null mutation causes neurological symptoms that mimic Rett syndrome. Nat Genet *27*, 322-326.

Hackenberg, M., Previti, C., Luque-Escamilla, P.L., Carpena, P., Martinez-Aroza, J., and Oliver, J.L. (2006). CpGcluster: a distance-based algorithm for CpG-island detection. BMC bioinformatics *7*, 446.

Hajkova, P., Ancelin, K., Waldmann, T., Lacoste, N., Lange, U.C., Cesari, F., Lee, C., Almouzni, G., Schneider, R., and Surani, M.A. (2008). Chromatin dynamics during epigenetic reprogramming in the mouse germ line. Nature *452*, 877-881.

Hajkova, P., Erhardt, S., Lane, N., Haaf, T., El-Maarri, O., Reik, W., Walter, J., and Surani, M.A. (2002). Epigenetic reprogramming in mouse primordial germ cells. Mechanisms of development *117*, 15-23.

Hansen, R.S., and Gartler, S.M. (1990). 5-Azacytidine-induced reactivation of the human X chromosome-linked PGK1 gene is associated with a large region of cytosine demethylation in the 5' CpG island. Proc Natl Acad Sci U S A *87*, 4174-4178.

Hansen, R.S., Stoger, R., Wijmenga, C., Stanek, A.M., Canfield, T.K., Luo, P., Matarazzo, M.R., D'Esposito, M., Feil, R., Gimelli, G., *et al.* (2000). Escape from gene silencing in ICF syndrome: evidence for advanced replication time as a major determinant. Hum Mol Genet *9*, 2575-2587.

Harrington, M.A., Jones, P.A., Imagawa, M., and Karin, M. (1988). Cytosine methylation does not affect binding of transcription factor Sp1. Proc Natl Acad Sci U S A *85*, 2066-2070.

Hata, K., Kusumi, M., Yokomine, T., Li, E., and Sasaki, H. (2006). Meiotic and epigenetic aberrations in Dnmt3L-deficient male germ cells. Molecular reproduction and development *73*, 116-122.

Hata, K., Okano, M., Lei, H., and Li, E. (2002). Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development *129*, 1983-1993.

Heard, E., Clerc, P., and Avner, P. (1997). X-chromosome inactivation in mammals. Annu Rev Genet *31*, 571-610.

Heard, E., Rougeulle, C., Arnaud, D., Avner, P., Allis, C.D., and Spector, D.L. (2001). Methylation of histone H3 at Lys-9 is an early mark on the X chromosome during X inactivation. Cell *107*, 727-738.

Heisler, L.E., Torti, D., Boutros, P.C., Watson, J., Chan, C., Winegarden, N., Takahashi, M., Yau, P., Huang, T.H., Farnham, P.J., *et al.* (2005). CpG Island microarray probe sequences derived from a physical library are representative of CpG Islands annotated on the human genome. Nucleic Acids Res *33*, 2952-2961.

Hellman, A., and Chess, A. (2007). Gene body-specific methylation on the active X chromosome. Science *315*, 1141-1143.

Hellmann-Blumberg, U., Hintz, M.F., Gatewood, J.M., and Schmid, C.W. (1993). Developmental differences in methylation of human Alu repeats. Mol Cell Biol *13*, 4523-4530.

Hendrich, B., and Bird, A. (1998). Identification and characterization of a family of mammalian methyl-CpG binding proteins. Mol Cell Biol *18*, 6538-6547.

Hendrich, B., and Bird, A. (2000). Mammalian methyltransferases and methyl-CpG-binding domains: proteins involved in DNA methylation. Curr Top Microbiol Immunol *249*, 55-74.

Hendrich, B., Guy, J., Ramsahoye, B., Wilson, V.A., and Bird, A. (2001). Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. Genes Dev *15*, 710-723.

Hendrich, B., Hardeland, U., Ng, H.H., Jiricny, J., and Bird, A. (1999). The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. Nature *401*, 301-304.

Herman, J.G., Graff, J.R., Myohanen, S., Nelkin, B.D., and Baylin, S.B. (1996). Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc Natl Acad Sci U S A *93*, 9821-9826.

Hermann, A., Gowher, H., and Jeltsch, A. (2004). Biochemistry and biology of mammalian DNA methyltransferases. Cell Mol Life Sci *61*, 2571-2587.

Hermann, A., Schmitt, S., and Jeltsch, A. (2003). The human Dnmt2 has residual DNA-(cytosine-C5) methyltransferase activity. J Biol Chem *278*, 31717-31721.

Hernandez-Munoz, I., Taghavi, P., Kuijl, C., Neefjes, J., and van Lohuizen, M. (2005). Association of BMI1 with polycomb bodies is dynamic and requires PRC2/EZH2 and the maintenance DNA methyltransferase DNMT1. Mol Cell Biol *25*, 11047-11058.

Ho, K.L., McNae, I.W., Schmiedeberg, L., Klose, R.J., Bird, A.P., and Walkinshaw, M.D. (2008). MeCP2 binding to DNA depends upon hydration at methyl-CpG. Mol Cell *29*, 525-531.

Honda, T., Tamura, G., Waki, T., Kawata, S., Terashima, M., Nishizuka, S., and Motoyama, T. (2004). Demethylation of MAGE promoters during gastric cancer progression. British journal of cancer *90*, 838-843.

Howlett, S.K., and Reik, W. (1991). Methylation levels of maternal and paternal genomes during preimplantation development. Development *113*, 119-127.

Huang, T.H., Perry, M.R., and Laux, D.E. (1999). Methylation profiling of CpG islands in human breast cancer cells. Hum Mol Genet *8*, 459-470.

Huntriss, J., Daniels, R., Bolton, V., and Monk, M. (1998). Imprinted expression of SNRPN in human preimplantation embryos. American journal of human genetics *63*, 1009-1014.

Illingworth, R., Kerr, A., Desousa, D., Jorgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J., *et al.* (2008). A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS biology *6*, e22.

Imamura, T., Ohgane, J., Ito, S., Ogawa, T., Hattori, N., Tanaka, S., and Shiota, K. (2001). CpG island of rat sphingosine kinase-1 gene: tissue-dependent DNA methylation status and multiple alternative first exons. Genomics *76*, 117-125.

Ioshikhes, I.P., Albert, I., Zanton, S.J., and Pugh, B.F. (2006). Nucleosome positions predicted through comparative genomics. Nat Genet *38*, 1210-1215.

Ioshikhes, I.P., and Zhang, M.Q. (2000). Large-scale human promoter mapping using CpG islands. Nat Genet *26*, 61-63.

Issa, J.P., Vertino, P.M., Boehm, C.D., Newsham, I.F., and Baylin, S.B. (1996). Switch from monoallelic to biallelic human IGF2 promoter methylation during aging and carcinogenesis. Proc Natl Acad Sci U S A *93*, 11757-11762.

Janicki, S.M., Tsukamoto, T., Salghetti, S.E., Tansey, W.P., Sachidanandam, R., Prasanth, K.V., Ried, T., Shav-Tal, Y., Bertrand, E., Singer, R.H., *et al.* (2004). From silencing to gene expression: real-time analysis in single cells. Cell *116*, 683-698.

Jeltsch, A. (2002). Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. Chembiochem *3*, 274-293.

Jia, D., Jurkowska, R.Z., Zhang, X., Jeltsch, A., and Cheng, X. (2007). Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. Nature *449*, 248-251.

Jones, P.A. (1999). The DNA methylation paradox. Trends Genet *15*, 34-37.

Jones, P.A. (2002). DNA methylation and cancer. Oncogene *21*, 5358-5360.

Jones, P.A., and Baylin, S.B. (2007). The epigenomics of cancer. Cell *128*, 683-692.

Jorgensen, H.F., Ben-Porath, I., and Bird, A.P. (2004). Mbd1 is recruited to both methylated and nonmethylated CpGs via distinct DNA binding domains. Mol Cell Biol *24*, 3387-3395.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res *110*, 462-467.

Kafri, T., Ariel, M., Brandeis, M., Shemer, R., Urven, L., McCarrey, J., Cedar, H., and Razin, A. (1992). Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. Genes Dev *6*, 705-714.

Kaji, K., Caballero, I.M., MacLeod, R., Nichols, J., Wilson, V.A., and Hendrich, B. (2006). The NuRD component Mbd3 is required for pluripotency of embryonic stem cells. Nat Cell Biol *8*, 285-292.

Kaji, K., Nichols, J., and Hendrich, B. (2007). Mbd3, a component of the NuRD co-repressor complex, is required for development of pluripotent cells. Development *134*, 1123-1132.

Kanellopoulou, C., Muljo, S.A., Kung, A.L., Ganesan, S., Drapkin, R., Jenuwein, T., Livingston, D.M., and Rajewsky, K. (2005). Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. Genes Dev *19*, 489-501.

Kangaspeska, S., Stride, B., Metivier, R., Polycarpou-Schwarz, M., Ibberson, D., Carmouche, R.P., Benes, V., Gannon, F., and Reid, G. (2008). Transient cyclical methylation of promoter DNA. Nature *452*, 112-115.

Karolchik, D., Kuhn, R.M., Baertsch, R., Barber, G.P., Clawson, H., Diekhans, M., Giardine, B., Harte, R.A., Hinrichs, A.S., Hsu, F., *et al.* (2008). The UCSC Genome Browser Database: 2008 update. Nucleic Acids Res *36*, D773-779.

Karpf, A.R., and Matsui, S. (2005). Genetic disruption of cytosine DNA methyltransferase enzymes induces chromosomal instability in human cancer cells. Cancer Res *65*, 8635-8639.

Kass, S.U., Landsberger, N., and Wolffe, A.P. (1997). DNA methylation directs a time-dependent repression of transcription initiation. Curr Biol *7*, 157-165.

Kelly, S.M., Pabit, S.A., Kitchen, C.M., Guo, P., Marfatia, K.A., Murphy, T.J., Corbett, A.H., and Berland, K.M. (2007). Recognition of polyadenosine RNA by zinc finger proteins. Proc Natl Acad Sci U S A *104*, 12306-12311.

Kennison, J.A. (1995). The Polycomb and trithorax group proteins of Drosophila: trans-regulators of homeotic gene function. Annu Rev Genet *29*, 289-303.

Keohane, A.M., O'Neill L, P., Belyaev, N.D., Lavender, J.S., and Turner, B.M. (1996). X-Inactivation and histone H4 acetylation in embryonic stem cells. Developmental biology *180*, 618-630.

Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., Segal, E., Pikarski, E., Young, R.A., Niveleau, A., Cedar, H., *et al.* (2006). Evidence for an instructive mechanism of de novo methylation in cancer cells. Nat Genet *38*, 149-153.

Kim, G.D., Ni, J., Kelesoglu, N., Roberts, R.J., and Pradhan, S. (2002). Co-operation and communication between the human maintenance and de novo DNA (cytosine-5) methyltransferases. Embo J *21*, 4183-4195.

Kitamura, E., Igarashi, J., Morohashi, A., Hida, N., Oinuma, T., Nemoto, N., Song, F., Ghosh, S., Held, W.A., Yoshida-Noro, C., *et al.* (2007). Analysis of tissue-specific differentially methylated regions (TDMs) in humans. Genomics *89*, 326-337.

Kleinjan, D.A., Seawright, A., Childs, A.J., and van Heyningen, V. (2004). Conserved elements in Pax6 intron 7 involved in (auto)regulation and alternative transcription. Developmental biology *265*, 462-477.

Klimasauskas, S., Kumar, S., Roberts, R.J., and Cheng, X. (1994). HhaI methyltransferase flips its target base out of the DNA helix. Cell *76*, 357-369.

Klose, R.J., and Bird, A.P. (2004). MeCP2 behaves as an elongated monomer that does not stably associate with the Sin3a chromatin remodeling complex. J Biol Chem *279*, 46490-46496.

Klose, R.J., and Bird, A.P. (2006). Genomic DNA methylation: the mark and its mediators. Trends Biochem Sci *31*, 89-97.

Klose, R.J., Sarraf, S.A., Schmiedeberg, L., McDermott, S.M., Stancheva, I., and Bird, A.P. (2005). DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. Mol Cell *19*, 667-678.

Klose, R.J., and Zhang, Y. (2007). Regulation of histone methylation by demethylimination and demethylation. Nat Rev Mol Cell Biol *8*, 307-318.

Knudson, A.G., Jr. (1971). Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A *68*, 820-823.

Kochanek, S., Renz, D., and Doerfler, W. (1993). DNA methylation in the Alu sequences of diploid and haploid primary human cells. Embo J *12*, 1141-1151.

Kodama, R., and Eguchi, G. (1994). Gene regulation and differentiation in vertebrate ocular tissues. Curr Opin Genet Dev *4*, 703-708.

Kondo, E., Gu, Z., Horii, A., and Fukushige, S. (2005). The thymine DNA glycosylase MBD4 represses transcription and is associated with methylated p16(INK4a) and hMLH1 genes. Mol Cell Biol *25*, 4388-4396.

Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693-705.

Koyama-Nasu, R., David, G., and Tanese, N. (2007). The F-box protein Fbl10 is a novel transcriptional repressor of c-Jun. Nat Cell Biol *9*, 1074-1080.

Kunert, N., Marhold, J., Stanke, J., Stach, D., and Lyko, F. (2003). A Dnmt2-like protein mediates DNA methylation in Drosophila. Development *130*, 5083-5090.

Kwabi-Addo, B., Chung, W., Shen, L., Ittmann, M., Wheeler, T., Jelinek, J., and Issa, J.P. (2007). Age-related DNA methylation changes in normal human prostate tissues. Clin Cancer Res *13*, 3796-3802.

Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., and Jenuwein, T. (2001). Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature *410*, 116-120.

Ladd-Acosta, C., Pevsner, J., Sabunciyan, S., Yolken, R.H., Webster, M.J., Dinkins, T., Callinan, P.A., Fan, J.B., Potash, J.B., and Feinberg, A.P. (2007). DNA methylation signatures within the human brain. American journal of human genetics *81*, 1304-1315.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. Genomics *13*, 1095-1107.

Le Guezennec, X., Vermeulen, M., Brinkman, A.B., Hoeijmakers, W.A., Cohen, A., Lasonder, E., and Stunnenberg, H.G. (2006). MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties. Mol Cell Biol *26*, 843-851.

Lee, J., Inoue, K., Ono, R., Ogonuki, N., Kohda, T., Kaneko-Ishino, T., Ogura, A., and Ishino, F. (2002). Erasing genomic imprinting memory in mouse clone embryos produced from day 11.5 primordial germ cells. Development *129*, 1807-1817.

Lee, J.H., Voo, K.S., and Skalnik, D.G. (2001). Identification and characterization of the DNA binding domain of CpG-binding protein. J Biol Chem *276*, 44669-44676.

Lee, M.G., Villa, R., Trojer, P., Norman, J., Yan, K.P., Reinberg, D., Di Croce, L., and Shiekhattar, R. (2007a). Demethylation of H3K27 regulates polycomb recruitment and H2A ubiquitination. Science *318*, 447-450.

Lee, M.K., Lynch, E.D., and King, M.C. (1998). SeqHelp: a program to analyze molecular sequences utilizing common computational resources. Genome Res *8*, 306-312.

Lee, N., Zhang, J., Klose, R.J., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2007b). The trithorax-group protein Lid is a histone H3 trimethyl-Lys4 demethylase. Nat Struct Mol Biol *14*, 341-343.

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., *et al.* (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell *125*, 301-313.

Lees-Murdock, D.J., De Felici, M., and Walsh, C.P. (2003). Methylation dynamics of repetitive DNA elements in the mouse germ cell lineage. Genomics *82*, 230-237.

Lehnertz, B., Ueda, Y., Derijck, A.A., Braunschweig, U., Perez-Burgos, L., Kubicek, S., Chen, T., Li, E., Jenuwein, T., and Peters, A.H. (2003). Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Curr Biol *13*, 1192-1200.

Lei, H., Oh, S.P., Okano, M., Juttermann, R., Goss, K.A., Jaenisch, R., and Li, E. (1996). De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. Development *122*, 3195-3205.

Leonhardt, H., Page, A.W., Weier, H.U., and Bestor, T.H. (1992). A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. Cell *71*, 865-873.

Lewis, J.D., Meehan, R.R., Henzel, W.J., Maurer-Fogy, I., Jeppesen, P., Klein, F., and Bird, A. (1992). Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. Cell *69*, 905-914.

Li, B., Carey, M., and Workman, J.L. (2007a). The role of chromatin during transcription. Cell *128*, 707-719.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. Nature *366*, 362-365.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915-926.

Li, H.G., Ding, X.F., Liao, A.H., Kong, X.B., and Xiong, C.L. (2007b). Expression of CatSper family transcripts in the mouse testis during post-natal development and human ejaculated spermatozoa: relationship to sperm motility. Mol Hum Reprod *13*, 299-306.

Liang, G., Chan, M.F., Tomigahara, Y., Tsai, Y.C., Gonzales, F.A., Li, E., Laird, P.W., and Jones, P.A. (2002). Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. Mol Cell Biol *22*, 480-491.

Lin, I.G., Han, L., Taghva, A., O'Brien, L.E., and Hsieh, C.L. (2002). Murine de novo methyltransferase Dnmt3a demonstrates strand asymmetry and site preference in the methylation of DNA in vitro. Mol Cell Biol *22*, 704-723.

Linhart, H.G., Lin, H., Yamada, Y., Moran, E., Steine, E.J., Gokhale, S., Lo, G., Cantu, E., Ehrich, M., He, T., *et al.* (2007). Dnmt3b promotes tumorigenesis in vivo by gene-specific de novo methylation and transcriptional silencing. Genes Dev *21*, 3110-3122.

Lippman, Z., and Martienssen, R. (2004). The role of RNA interference in heterochromatic silencing. Nature *431*, 364-370.

Lock, L.F., Takagi, N., and Martin, G.R. (1987). Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. Cell *48*, 39-46.

Lorincz, M.C., Dickerson, D.R., Schmitt, M., and Groudine, M. (2004). Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. Nat Struct Mol Biol *11*, 1068-1075.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251-260.

Luikenhuis, S., Wutz, A., and Jaenisch, R. (2001). Antisense transcription through the Xist locus mediates Tsix function in embryonic stem cells. Mol Cell Biol *21*, 8512-8520.

Lyko, F., Ramsahoye, B.H., and Jaenisch, R. (2000). DNA methylation in Drosophila melanogaster. Nature *408*, 538-540.

Lyst, M.J., Nan, X., and Stancheva, I. (2006). Regulation of MBD1-mediated transcriptional repression by SUMO and PIAS proteins. Embo J *25*, 5317-5328.

Ma, Q., Alder, H., Nelson, K.K., Chatterjee, D., Gu, Y., Nakamura, T., Canaani, E., Croce, C.M., Siracusa, L.D., and Buchberg, A.M. (1993). Analysis of the murine All-1 gene reveals conserved domains with human ALL-1 and identifies a motif shared with DNA methyltransferases. Proc Natl Acad Sci U S A *90*, 6350-6354.

Maatouk, D.M., Kellam, L.D., Mann, M.R., Lei, H., Li, E., Bartolomei, M.S., and Resnick, J.L. (2006). DNA methylation is a primary mechanism for silencing postmigratory primordial germ cell genes in both germ cell and somatic cell lineages. Development *133*, 3411-3418.

Macleod, D., Ali, R.R., and Bird, A. (1998). An alternative promoter in the mouse major histocompatibility complex class II I-Abeta gene: implications for the origin of CpG islands. Mol Cell Biol *18*, 4433-4443.

Macleod, D., Charlton, J., Mullins, J., and Bird, A.P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev *8*, 2282-2292.

Maloisel, L., and Rossignol, J.L. (1998). Suppression of crossing-over by DNA methylation in Ascobolus. Genes Dev *12*, 1381-1389.

Mao, D.Y., Watson, J.D., Yan, P.S., Barsyte-Lovejoy, D., Khosravi, F., Wong, W.W., Farnham, P.J., Huang, T.H., and Penn, L.Z. (2003). Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression. Curr Biol *13*, 882-886.

Margueron, R., Trojer, P., and Reinberg, D. (2005). The key to development: interpreting the histone code? Curr Opin Genet Dev *15*, 163-176.

Martin, C., and Zhang, Y. (2007). Mechanisms of epigenetic inheritance. Current opinion in cell biology *19*, 266-272.

Matsuyama, T., Kimura, M.T., Koike, K., Abe, T., Nakano, T., Asami, T., Ebisuzaki, T., Held, W.A., Yoshida, S., and Nagase, H. (2003). Global methylation screening in the Arabidopsis thaliana and Mus musculus genome: applications of virtual image restriction landmark genomic scanning (Vi-RLGS). Nucleic Acids Res *31*, 4490-4496.

Matzke, M.A., and Birchler, J.A. (2005). RNAi-mediated pathways in the nucleus. Nature reviews *6*, 24-35.

Mayer, W., Niveleau, A., Walter, J., Fundele, R., and Haaf, T. (2000). Demethylation of the zygotic paternal genome. Nature *403*, 501-502.

McQueen, H.A., Clark, V.H., Bird, A.P., Yerle, M., and Archibald, A.L. (1997). CpG islands of the pig. Genome Res *7*, 924-931.

McQueen, H.A., Fantes, J., Cross, S.H., Clark, V.H., Archibald, A.L., and Bird, A.P. (1996). CpG islands of chicken are concentrated on microchromosomes. Nat Genet *12*, 321-324.

Meehan, R.R., Lewis, J.D., McKay, S., Kleiner, E.L., and Bird, A.P. (1989). Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. Cell *58*, 499-507.

Mermoud, J.E., Popova, B., Peters, A.H., Jenuwein, T., and Brockdorff, N. (2002). Histone H3 lysine 9 methylation occurs rapidly at the onset of random X chromosome inactivation. Curr Biol *12*, 247-251.

Metivier, R., Gallais, R., Tiffoche, C., Le Peron, C., Jurkowska, R.Z., Carmouche, R.P., Ibberson, D., Barath, P., Demay, F., Reid, G., *et al.* (2008). Cyclical DNA methylation of a transcriptionally active promoter. Nature *452*, 45-50.

Meunier, J., Khelifi, A., Navratil, V., and Duret, L. (2005). Homology-dependent methylation in primate repetitive DNA. Proc Natl Acad Sci U S A *102*, 5471-5476.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., *et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Millar, C.B., Guy, J., Sansom, O.J., Selfridge, J., MacDougall, E., Hendrich, B., Keightley, P.D., Bishop, S.M., Clarke, A.R., and Bird, A. (2002). Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. Science *297*, 403-405.

Miniou, P., Bourc'his, D., Molina Gomes, D., Jeanpierre, M., and Viegas-Pequignot, E. (1997). Undermethylation of Alu sequences in ICF syndrome: molecular and in situ analysis. Cytogenetics and cell genetics *77*, 308-313.

Mizuguchi, G., Shen, X., Landry, J., Wu, W.H., Sen, S., and Wu, C. (2004). ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. Science *303*, 343-348.

Mohandas, T., Sparkes, R.S., and Shapiro, L.J. (1981). Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. Science *211*, 393-396.

Monk, M., Adams, R.L., and Rinaldi, A. (1991). Decrease in DNA methylase activity during preimplantation development in the mouse. Development *112*, 189-192.

Monk, M., Boubelik, M., and Lehnert, S. (1987). Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development *99*, 371-382.

Morales-Ruiz, T., Ortega-Galisteo, A.P., Ponferrada-Marin, M.I., Martinez-Macias, M.I., Ariza, R.R., and Roldan-Arjona, T. (2006). DEMETER and REPRESSOR OF SILENCING 1 encode 5-methylcytosine DNA glycosylases. Proc Natl Acad Sci U S A *103*, 6853-6858.

Morris, K.V., Chan, S.W., Jacobsen, S.E., and Looney, D.J. (2004). Small interfering RNA-induced transcriptional gene silencing in human cells. Science *305*, 1289-1292.

Muchardt, C., and Yaniv, M. (1999). ATP-dependent chromatin remodelling: SWI/SNF and Co. are on the job. J Mol Biol *293*, 187-198.

Myant, K., and Stancheva, I. (2008). LSH cooperates with DNA methyltransferases to repress transcription. Mol Cell Biol *28*, 215-226.

Nakamura, T., Arai, Y., Umehara, H., Masuhara, M., Kimura, T., Taniguchi, H., Sekimoto, T., Ikawa, M., Yoneda, Y., Okabe, M., *et al.* (2007). PGC7/Stella protects against DNA demethylation in early embryogenesis. Nat Cell Biol *9*, 64-71.

Nakamura, T., Mori, T., Tada, S., Krajewski, W., Rozovskaia, T., Wassell, R., Dubois, G., Mazo, A., Croce, C.M., and Canaani, E. (2002). ALL-1 is a histone methyltransferase that

assembles a supercomplex of proteins involved in transcriptional regulation. Mol Cell *10*, 1119-1128.

Nan, X., and Bird, A. (2001). The biological functions of the methyl-CpG-binding protein MeCP2 and its implication in Rett syndrome. Brain Dev *23 Suppl 1*, S32-37.

Nan, X., Campoy, F.J., and Bird, A. (1997). MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. Cell *88*, 471-481.

Nan, X., Meehan, R.R., and Bird, A. (1993). Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. Nucleic Acids Res *21*, 4886-4892.

Nan, X., Ng, H.H., Johnson, C.A., Laherty, C.D., Turner, B.M., Eisenman, R.N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386-389.

Ng, H.H., Zhang, Y., Hendrich, B., Johnson, C.A., Turner, B.M., Erdjument-Bromage, H., Tempst, P., Reinberg, D., and Bird, A. (1999). MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. Nat Genet *23*, 58-61.

Nguyen, C., Liang, G., Nguyen, T.T., Tsao-Wei, D., Groshen, S., Lubbert, M., Zhou, J.H., Benedict, W.F., and Jones, P.A. (2001). Susceptibility of nonpromoter CpG islands to de novo methylation in normal and neoplastic cells. J Natl Cancer Inst *93*, 1465-1472.

Nguyen, D.K., and Disteche, C.M. (2006). Dosage compensation of the active X chromosome in mammals. Nat Genet *38*, 47-53.

Nightingale, K.P., O'Neill, L.P., and Turner, B.M. (2006). Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. Curr Opin Genet Dev *16*, 125-136.

Nikitina, T., Shi, X., Ghosh, R.P., Horowitz-Scherer, R.A., Hansen, J.C., and Woodcock, C.L. (2007). Multiple modes of interaction between the methylated DNA binding protein MeCP2 and chromatin. Mol Cell Biol *27*, 864-877.

Norris, D.P., Brockdorff, N., and Rastan, S. (1991). Methylation status of CpG-rich islands on active and inactive mouse X chromosomes. Mamm Genome *1*, 78-83.

Norris, D.P., Patel, D., Kay, G.F., Penny, G.D., Brockdorff, N., Sheardown, S.A., and Rastan, S. (1994). Evidence that random and imprinted Xist expression is controlled by preemptive methylation. Cell *77*, 41-51.

Oakes, C.C., La Salle, S., Smiraglia, D.J., Robaire, B., and Trasler, J.M. (2007). A unique configuration of genome-wide DNA methylation patterns in the testis. Proc Natl Acad Sci U S A *104*, 228-233.

Oakes, C.C., Smiraglia, D.J., Plass, C., Trasler, J.M., and Robaire, B. (2003). Aging results in hypermethylation of ribosomal DNA in sperm and liver of male rats. Proc Natl Acad Sci U S A *100*, 1775-1780.

Oda, M., Yamagiwa, A., Yamamoto, S., Nakayama, T., Tsumura, A., Sasaki, H., Nakao, K., Li, E., and Okano, M. (2006). DNA methylation regulates long-range gene silencing of an X-linked homeobox gene cluster in a lineage-specific manner. Genes Dev *20*, 3382-3394.

Ohki, I., Shimotake, N., Fujita, N., Jee, J., Ikegami, T., Nakao, M., and Shirakawa, M. (2001). Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. Cell *105*, 487-497.

Ohki, I., Shimotake, N., Fujita, N., Nakao, M., and Shirakawa, M. (1999). Solution structure of the methyl-CpG-binding domain of the methylation-dependent transcriptional repressor MBD1. Embo J *18*, 6653-6661.

Ohm, J.E., and Baylin, S.B. (2007). Stem cell chromatin patterns: an instructive mechanism for DNA hypermethylation? Cell cycle (Georgetown, Tex *6*, 1040-1043.

Ohm, J.E., McGarvey, K.M., Yu, X., Cheng, L., Schuebel, K.E., Cope, L., Mohammad, H.P., Chen, W., Daniel, V.C., Yu, W., *et al.* (2007). A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat Genet *39*, 237-242.

Okamoto, I., Otte, A.P., Allis, C.D., Reinberg, D., and Heard, E. (2004). Epigenetic dynamics of imprinted X inactivation during early mouse development. Science *303*, 644-649.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247-257.

Okano, M., Xie, S., and Li, E. (1998a). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. Nat Genet *19*, 219-220.

Okano, M., Xie, S., and Li, E. (1998b). Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells. Nucleic Acids Res *26*, 2536-2540.

Ooi, S.K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.P., Allis, C.D., *et al.* (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature *448*, 714-717.

Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R., Dean, W., Reik, W., and Walter, J. (2000). Active demethylation of the paternal genome in the mouse zygote. Curr Biol *10*, 475-478.

Panning, B., and Jaenisch, R. (1996). DNA hypomethylation can activate Xist expression and silence X-linked genes. Genes Dev *10*, 1991-2002.

Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., and Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. Nature *379*, 131-137.

Penterman, J., Zilberman, D., Huh, J.H., Ballinger, T., Henikoff, S., and Fischer, R.L. (2007). DNA demethylation in the Arabidopsis genome. Proc Natl Acad Sci U S A *104*, 6752-6757.

Phi-van, L., and Stratling, W.H. (1999). An origin of bidirectional DNA replication is located within a CpG island at the 3" end of the chicken lysozyme gene. Nucleic Acids Res *27*, 3009-3017.

Pinto, M., Wu, Y., Mensink, R.G., Cirnes, L., Seruca, R., and Hofstra, R.M. (2008). Somatic mutations in mismatch repair genes in sporadic gastric carcinomas are not a cause but a consequence of the mutator phenotype. Cancer Genet Cytogenet *180*, 110-114.

Plass, C., Yu, F., Yu, L., Strout, M.P., El-Rifai, W., Elonen, E., Knuutila, S., Marcucci, G., Young, D.C., Held, W.A., *et al.* (1999). Restriction landmark genome scanning for aberrant methylation in primary refractory and relapsed acute myeloid leukemia; involvement of the WIT-1 gene. Oncogene *18*, 3159-3165.

Plath, K., Fang, J., Mlynarczyk-Evans, S.K., Cao, R., Worringer, K.A., Wang, H., de la Cruz, C.C., Otte, A.P., Panning, B., and Zhang, Y. (2003). Role of histone H3 lysine 27 methylation in X inactivation. Science *300*, 131-135.

Pollack, Y., Stein, R., Razin, A., and Cedar, H. (1980). Methylation of foreign DNA sequences in eukaryotic cells. Proc Natl Acad Sci U S A *77*, 6463-6467.

Ponger, L., Duret, L., and Mouchiroud, D. (2001). Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res *11*, 1854-1860.

Ponger, L., and Mouchiroud, D. (2002). CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics *18*, 631-633.

Prestridge, D.S., and Burks, C. (1993). The density of transcriptional elements in promoter and non-promoter sequences. Hum Mol Genet *2*, 1449-1453.

Priest, J.H., Heady, J.E., and Priest, R.E. (1967). Delayed onset of replication of human X chromosomes. The Journal of cell biology *35*, 483-487.

Proffitt, J.H., Davie, J.R., Swinton, D., and Hattman, S. (1984). 5-Methylcytosine is not detectable in Saccharomyces cerevisiae DNA. Mol Cell Biol *4*, 985-988.

Prokhortchouk, A., Hendrich, B., Jorgensen, H., Ruzov, A., Wilm, M., Georgiev, G., Bird, A., and Prokhortchouk, E. (2001). The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. Genes Dev *15*, 1613-1618.

Ptashne, M. (2005). Regulation of transcription: from lambda to eukaryotes. Trends Biochem Sci *30*, 275-279.

Qiu, C., Sawada, K., Zhang, X., and Cheng, X. (2002). The PWWP domain of mammalian DNA methyltransferase Dnmt3b defines a new family of DNA-binding folds. Nature structural biology *9*, 217-224.

Ramsahoye, B.H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A.P., and Jaenisch, R. (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc Natl Acad Sci U S A *97*, 5237-5242.

Ratnam, S., Mertineit, C., Ding, F., Howell, C.Y., Clarke, H.J., Bestor, T.H., Chaillet, J.R., and Trasler, J.M. (2002). Dynamics of Dnmt1 methyltransferase expression and intracellular localization during oogenesis and preimplantation development. Developmental biology *245*, 304-314.

Rauch, T., Li, H., Wu, X., and Pfeifer, G.P. (2006). MIRA-assisted microarray analysis, a new technology for the determination of DNA methylation patterns, identifies frequent methylation of homeodomain-containing genes in lung cancer cells. Cancer Res *66*, 7939-7947.

Redon, C., Pilch, D., Rogakou, E., Sedelnikova, O., Newrock, K., and Bonner, W. (2002). Histone H2A variants H2AX and H2AZ. Curr Opin Genet Dev *12*, 162-169.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. Nature *447*, 425-432.

Reik, W., Dean, W., and Walter, J. (2001). Epigenetic reprogramming in mammalian development. Science *293*, 1089-1093.

Rein, T., Zorbas, H., and DePamphilis, M.L. (1997). Active mammalian replication origins are associated with a high-density cluster of mCpG dinucleotides. Mol Cell Biol *17*, 416-426.

Reinisch, K.M., Chen, L., Verdine, G.L., and Lipscomb, W.N. (1995). The crystal structure of HaeIII methyltransferase convalently complexed to DNA: an extrahelical cytosine and rearranged base pairing. Cell *82*, 143-153.

Riccio, A., Aaltonen, L.A., Godwin, A.K., Loukola, A., Percesepe, A., Salovaara, R., Masciullo, V., Genuardi, M., Paravatou-Petsotas, M., Bassi, D.E., *et al.* (1999). The DNA repair gene MBD4 (MED1) is mutated in human carcinomas with microsatellite instability. Nat Genet *23*, 266-268.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet *16*, 276-277.

Riethman, H., Ambrosini, A., Castaneda, C., Finklestein, J., Hu, X.L., Mudunuri, U., Paul, S., and Wei, J. (2004). Mapping and initial analysis of human subtelomeric sequence assemblies. Genome Res *14*, 18-28.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., *et al.* (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311-1323.

Robertson, A.K., Geiman, T.M., Sankpal, U.T., Hager, G.L., and Robertson, K.D. (2004). Effects of chromatin structure on the enzymatic and DNA binding functions of DNA methyltransferases DNMT1 and Dnmt3a in vitro. Biochem Biophys Res Commun *322*, 110-118.

Robertson, K.D., Ait-Si-Ali, S., Yokochi, T., Wade, P.A., Jones, P.L., and Wolffe, A.P. (2000). DNMT1 forms a complex with Rb, E2F1 and HDAC1 and represses transcription from E2F-responsive promoters. Nat Genet *25*, 338-342.

Rodriguez, J., Frigola, J., Vendrell, E., Risques, R.A., Fraga, M.F., Morales, C., Moreno, V., Esteller, M., Capella, G., Ribas, M., *et al.* (2006). Chromosomal instability correlates with genome-wide DNA demethylation in human primary colorectal cancers. Cancer Res *66*, 8462-9468.

Roh, T.Y., Cuddapah, S., and Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. Genes Dev *19*, 542-552.

Roll, J.D., Rivenbark, A.G., Jones, W.D., and Coleman, W.B. (2008). DNMT3b overexpression contributes to a hypermethylator phenotype in human breast cancer cell lines. Molecular cancer *7*, 15.

Rollins, R.A., Haghighi, F., Edwards, J.R., Das, R., Zhang, M.Q., Ju, J., and Bestor, T.H. (2006). Large-scale structure of genomic methylation patterns. Genome Res *16*, 157-163.

Rountree, M.R., Bachman, K.E., and Baylin, S.B. (2000). DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci. Nat Genet *25*, 269-277.

Rubin, C.M., VandeVoort, C.A., Teplitz, R.L., and Schmid, C.W. (1994). Alu repeated DNAs are differentially methylated in primate germ cells. Nucleic Acids Res *22*, 5121-5127.

Ruzov, A., Dunican, D.S., Prokhortchouk, A., Pennings, S., Stancheva, I., Prokhortchouk, E., and Meehan, R.R. (2004). Kaiso is a genome-wide repressor of transcription that is essential for amphibian development. Development *131*, 6185-6194.

Sado, T., Li, E., and Sasaki, H. (2002). Effect of TSIX disruption on XIST expression in male ES cells. Cytogenet Genome Res *99*, 115-118.

Saha, A., Wittmeyer, J., and Cairns, B.R. (2006). Chromatin remodelling: the industrial revolution of DNA around histones. Nat Rev Mol Cell Biol *7*, 437-447.

Sakai, Y., Suetake, I., Shinozaki, F., Yamashina, S., and Tajima, S. (2004). Co-expression of de novo DNA methyltransferases Dnmt3a2 and Dnmt3L in gonocytes of mouse embryos. Gene Expr Patterns *5*, 231-237.

Sakamoto, Y., Watanabe, S., Ichimura, T., Kawasuji, M., Koseki, H., Baba, H., and Nakao, M. (2007). Overlapping roles of the methylated DNA-binding protein MBD1 and polycomb group proteins in transcriptional repression of HOXA genes and heterochromatin foci formation. J Biol Chem *282*, 16391-16400.

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume, D.A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nature reviews *8*, 424-436.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A *74*, 5463-5467.

Santoro, R., and Grummt, I. (2005). Epigenetic mechanism of rRNA gene silencing: temporal order of NoRC-mediated histone modification, chromatin remodeling, and DNA methylation. Mol Cell Biol *25*, 2539-2546.

Santoro, R., Li, J., and Grummt, I. (2002). The nucleolar remodeling complex NoRC mediates heterochromatin formation and silencing of ribosomal gene transcription. Nat Genet *32*, 393-396.

Santos, F., Hendrich, B., Reik, W., and Dean, W. (2002). Dynamic reprogramming of DNA methylation in the early mouse embryo. Developmental biology *241*, 172-182.

Sarraf, S.A., and Stancheva, I. (2004). Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly. Mol Cell *15*, 595-605.

Satchwell, S.C., Drew, H.R., and Travers, A.A. (1986). Sequence periodicities in chicken nucleosome core DNA. J Mol Biol *191*, 659-675.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A *103*, 1412-1417.

Schilling, E., and Rehli, M. (2007). Global, comparative analysis of tissue-specific promoter CpG methylation. Genomics.

Schlesinger, Y., Straussman, R., Keshet, I., Farkash, S., Hecht, M., Zimmerman, J., Eden, E., Yakhini, Z., Ben-Shushan, E., Reubinoff, B.E., *et al.* (2007). Polycomb-mediated

methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat Genet *39*, 232-236.

Schmid, C.W. (1996). Alu: structure, origin, evolution, significance and function of one-tenth of human DNA. Progress in nucleic acid research and molecular biology *53*, 283-319.

Schoeftner, S., Sengupta, A.K., Kubicek, S., Mechtler, K., Spahn, L., Koseki, H., Jenuwein, T., and Wutz, A. (2006). Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. Embo J *25*, 3110-3122.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887-898.

Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B., and Cavalli, G. (2007). Genome regulation by polycomb and trithorax proteins. Cell *128*, 735-745.

Schumacher, A., Kapranov, P., Kaminsky, Z., Flanagan, J., Assadzadeh, A., Yau, P., Virtanen, C., Winegarden, N., Cheng, J., Gingeras, T., *et al.* (2006). Microarray-based DNA methylation profiling: technology and applications. Nucleic Acids Res *34*, 528-542.

Selker, E.U. (2002). Repeat-induced gene silencing in fungi. Advances in genetics *46*, 439-450.

Selker, E.U., Tountas, N.A., Cross, S.H., Margolin, B.S., Murphy, J.G., Bird, A.P., and Freitag, M. (2003). The methylated component of the Neurospora crassa genome. Nature *422*, 893-897.

Shao, Z., Raible, F., Mollaaghababa, R., Guyon, J.R., Wu, C.T., Bender, W., and Kingston, R.E. (1999). Stabilization of chromatin structure by PRC1, a Polycomb complex. Cell *98*, 37-46.

Shawi, M., and Autexier, C. (2008). Telomerase, senescence and ageing. Mechanisms of ageing and development *129*, 3-10.

Sheardown, S.A., Duthie, S.M., Johnston, C.M., Newall, A.E., Formstone, E.J., Arkell, R.M., Nesterova, T.B., Alghisi, G.C., Rastan, S., and Brockdorff, N. (1997). Stabilization of Xist RNA mediates initiation of X chromosome inactivation. Cell *91*, 99-107.

Shen, L., Kondo, Y., Guo, Y., Zhang, J., Zhang, L., Ahmed, S., Shu, J., Chen, X., Waterland, R.A., and Issa, J.P. (2007). Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. PLoS Genet *3*, 2023-2036.

Shi, H., Yan, P.S., Chen, C.M., Rahmatpanah, F., Lofton-Day, C., Caldwell, C.W., and Huang, T.H. (2002). Expressed CpG island sequence tag microarray for dual screening of DNA hypermethylation and gene silencing in cancer cells. Cancer Res *62*, 3214-3220.

Shi, Y., Lan, F., Matson, C., Mulligan, P., Whetstine, J.R., Cole, P.A., Casero, R.A., and Shi, Y. (2004). Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell *119*, 941-953.

Shiota, K. (2004). DNA methylation profiles of CpG islands for cellular differentiation and development in mammals. Cytogenet Genome Res *105*, 325-334.

Shiota, K., Kogo, Y., Ohgane, J., Imamura, T., Urano, A., Nishino, K., Tanaka, S., and Hattori, N. (2002). Epigenetic marks by DNA methylation specific to stem, germ and somatic cells in mice. Genes Cells *7*, 961-969.

Shiraishi, M., Chuu, Y.H., and Sekiya, T. (1999). Isolation of DNA fragments associated with methylated CpG islands in human adenocarcinomas of the lung using a methylated DNA binding column and denaturing gradient gel electrophoresis. Proc Natl Acad Sci U S A *96*, 2913-2918.

Shiraishi, M., Lerman, L.S., and Sekiya, T. (1995). Preferential isolation of DNA fragments associated with CpG islands. Proc Natl Acad Sci U S A *92*, 4229-4233.

Shiraishi, M., Sekiguchi, A., Oates, A.J., Terry, M.J., and Miyamoto, Y. (2002). HOX gene clusters are hotspots of de novo methylation in CpG islands of human lung adenocarcinomas. Oncogene *21*, 3659-3662.

Shiraishi, M., Sekiguchi, A., Oates, A.J., Terry, M.J., Miyamoto, Y., and Sekiya, T. (2004). Methyl-CpG binding domain column chromatography as a tool for the analysis of genomic DNA methylation. Anal Biochem *329*, 1-10.

Silva, J., Mak, W., Zvetkova, I., Appanah, R., Nesterova, T.B., Webster, Z., Peters, A.H., Jenuwein, T., Otte, A.P., and Brockdorff, N. (2003). Establishment of histone h3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. Developmental cell *4*, 481-495.

Simmen, M.W., Leitgeb, S., Charlton, J., Jones, S.J., Harris, B.R., Clark, V.H., and Bird, A. (1999). Nonmethylated transposable elements and methylated genes in a chordate genome. Science *283*, 1164-1167.

Sinkkonen, L., Hugenschmidt, T., Berninger, P., Gaidatzis, D., Mohn, F., Artus-Revel, C.G., Zavolan, M., Svoboda, P., and Filipowicz, W. (2008). MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. Nat Struct Mol Biol *15*, 259-267.

Sleutels, F., Zwart, R., and Barlow, D.P. (2002). The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature *415*, 810-813.

Smith, M.M. (2002). Centromeres and variant histones: what, where, when and why? Current opinion in cell biology *14*, 279-285.

Smyth, G.K., and Speed, T. (2003). Normalization of cDNA microarray data. Methods (San Diego, Calif *31*, 265-273.

Song, F., Smith, J.F., Kimura, M.T., Morrow, A.D., Matsuyama, T., Nagase, H., and Held, W.A. (2005). Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. Proc Natl Acad Sci U S A *102*, 3336-3341.

Stein, R., Razin, A., and Cedar, H. (1982). In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. Proc Natl Acad Sci U S A *79*, 3418-3422.

Strichman-Almashanu, L.Z., Lee, R.S., Onyango, P.O., Perlman, E., Flam, F., Frieman, M.B., and Feinberg, A.P. (2002). A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. Genome Res *12*, 543-554.

Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., *et al.* (2002). Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A *99*, 4465-4470.

Sudarsanam, P., Iyer, V.R., Brown, P.O., and Winston, F. (2000). Whole-genome expression analysis of snf/swi mutants of Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *97*, 3364-3369.

Suetake, I., Miyazaki, J., Murakami, C., Takeshima, H., and Tajima, S. (2003). Distinct enzymatic properties of recombinant mouse DNA methyltransferases Dnmt3a and Dnmt3b. Journal of biochemistry *133*, 737-744.

Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H., and Tajima, S. (2004). DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. J Biol Chem *279*, 27816-27823.

Sun, L.Q., Lee, D.W., Zhang, Q., Xiao, W., Raabe, E.H., Meeker, A., Miao, D., Huso, D.L., and Arceci, R.J. (2004). Growth retardation and premature aging phenotypes in mice with disruption of the SNF2-like gene, PASG. Genes Dev *18*, 1035-1046.

Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nature reviews.

Suzuki, M.M., Kerr, A.R., De Sousa, D., and Bird, A. (2007). CpG methylation is targeted to transcription units in an invertebrate genome. Genome Res *17*, 625-631.

Szabo, P., Tang, S.H., Rentsendorj, A., Pfeifer, G.P., and Mann, J.R. (2000). Maternal-specific footprints at putative CTCF sites in the H19 imprinting control region give evidence for insulator function. Curr Biol *10*, 607-610.

Takai, D., and Jones, P.A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A *99*, 3740-3745.

Tamaru, H., Zhang, X., McMillen, D., Singh, P.B., Nakayama, J., Grewal, S.I., Allis, C.D., Cheng, X., and Selker, E.U. (2003). Trimethylated lysine 9 of histone H3 is a mark for DNA methylation in Neurospora crassa. Nat Genet *34*, 75-79.

Tazi, J., and Bird, A. (1990). Alternative chromatin structure at CpG islands. Cell *60*, 909-920.

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. Genome Res *13*, 2129-2141.

Ting, A.H., McGarvey, K.M., and Baylin, S.B. (2006). The cancer epigenome--components and functional correlates. Genes Dev *20*, 3215-3231.

Ting, A.H., Schuebel, K.E., Herman, J.G., and Baylin, S.B. (2005). Short double-stranded RNA induces transcriptional gene silencing in human cancer cells in the absence of DNA methylation. Nat Genet *37*, 906-910.

Ting, A.H., Suzuki, H., Cope, L., Schuebel, K.E., Lee, B.H., Toyota, M., Imai, K., Shinomura, Y., Tokino, T., and Baylin, S.B. (2008). A requirement for DICER to maintain full promoter CpG island hypermethylation in human cancer cells. Cancer Res *68*, 2570-2575.

Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J.G., Baylin, S.B., and Issa, J.P. (1999). CpG island methylator phenotype in colorectal cancer. Proc Natl Acad Sci U S A *96*, 8681-8686.

Tremblay, K.D., Duran, K.L., and Bartolomei, M.S. (1997). A 5' 2-kilobase-pair region of the imprinted mouse H19 gene exhibits exclusive paternal methylation throughout development. Mol Cell Biol *17*, 4322-4329.

Tribioli, C., Tamanini, F., Patrosso, C., Milanesi, L., Villa, A., Pergolizzi, R., Maestrini, E., Rivella, S., Bione, S., Mancini, M., *et al.* (1992). Methylation and sequence analysis around EagI sites: identification of 28 new CpG islands in XQ24-XQ28. Nucleic Acids Res *20*, 727-733.

Tudor, M., Akbarian, S., Chen, R.Z., and Jaenisch, R. (2002). Transcriptional profiling of a mouse model for Rett syndrome reveals subtle transcriptional changes in the brain. Proc Natl Acad Sci U S A *99*, 15536-15541.

Tweedie, S., Charlton, J., Clark, V., and Bird, A. (1997). Methylation of genomes and genes at the invertebrate-vertebrate boundary. Mol Cell Biol *17*, 1469-1475.

Ueda, T., Abe, K., Miura, A., Yuzuriha, M., Zubair, M., Noguchi, M., Niwa, K., Kawase, Y., Kono, T., Matsuda, Y., *et al.* (2000). The paternal methylation imprint of the mouse H19 locus is acquired in the gonocyte stage during foetal testis development. Genes Cells *5*, 649-659.

Ueki, T., Toyota, M., Skinner, H., Walter, K.M., Yeo, C.J., Issa, J.P., Hruban, R.H., and Goggins, M. (2001). Identification and characterization of differentially methylated CpG islands in pancreatic carcinoma. Cancer Res *61*, 8540-8546.

Ushijima, T., Watanabe, N., Shimizu, K., Miyamoto, K., Sugimura, T., and Kaneda, A. (2005). Decreased fidelity in replicating CpG methylation patterns in cancer cells. Cancer Res *65*, 11-17.

van Raamsdonk, C.D., and Tilghman, S.M. (2000). Dosage requirement and allelic expression of PAX6 during lens placode formation. Development *127*, 5439-5448.

Vardimon, L., Kressmann, A., Cedar, H., Maechler, M., and Doerfler, W. (1982). Expression of a cloned adenovirus gene is inhibited by in vitro methylation. Proc Natl Acad Sci U S A *79*, 1073-1077.

Varga-Weisz, P. (2001). ATP-dependent chromatin remodeling factors: nucleosome shufflers with many missions. Oncogene *20*, 3076-3085.

Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.W., van Schaik, F.M., Varier, R.A., Baltissen, M.P., Stunnenberg, H.G., Mann, M., and Timmers, H.T. (2007). Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. Cell *131*, 58-69.

Viegas-Pequignot, E., Dutrillaux, B., and Thomas, G. (1988). Inactive X chromosome has the highest concentration of unmethylated Hha I sites. Proc Natl Acad Sci U S A *85*, 7657-7660.

Vire, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.M., *et al.* (2006). The Polycomb group protein EZH2 directly controls DNA methylation. Nature *439*, 871-874.

Voo, K.S., Carlone, D.L., Jacobsen, B.M., Flodin, A., and Skalnik, D.G. (2000). Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. Mol Cell Biol *20*, 2108-2121.

Wakefield, R.I., Smith, B.O., Nan, X., Free, A., Soteriou, A., Uhrin, D., Bird, A.P., and Barlow, P.N. (1999). The solution structure of the domain from MeCP2 that binds to methylated DNA. J Mol Biol *291*, 1055-1065.

Walsh, C.P., Chaillet, J.R., and Bestor, T.H. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat Genet *20*, 116-117.

Watanabe, T., Inoue, S., Hiroi, H., Orimo, A., Kawashima, H., and Muramatsu, M. (1998). Isolation of estrogen-responsive genes with a CpG island library. Mol Cell Biol *18*, 442-449.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Watt, F., and Molloy, P.L. (1988). Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. Genes Dev *2*, 1136-1143.

Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L., and Schubeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M., and Schubeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet *39*, 457-466.

Weber, M., and Schubeler, D. (2007). Genomic patterns of DNA methylation: targets and function of an epigenetic mark. Current opinion in cell biology *19*, 273-280.

Webster, K.E., O'Bryan, M.K., Fletcher, S., Crewther, P.E., Aapola, U., Craig, J., Harrison, D.K., Aung, H., Phutikanit, N., Lyle, R., *et al.* (2005). Meiotic and epigenetic defects in Dnmt3L-knockout mouse spermatogenesis. Proc Natl Acad Sci U S A *102*, 4068-4073.

Weinmann, A.S., Yan, P.S., Oberley, M.J., Huang, T.H., and Farnham, P.J. (2002). Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev *16*, 235-244.

Weisenberger, D.J., Campan, M., Long, T.I., Kim, M., Woods, C., Fiala, E., Ehrlich, M., and Laird, P.W. (2005). Analysis of repetitive element DNA methylation by MethyLight. Nucleic Acids Res *33*, 6823-6836.

Widschwendter, M., Fiegl, H., Egle, D., Mueller-Holzner, E., Spizzo, G., Marth, C., Weisenberger, D.J., Campan, M., Young, J., Jacobs, I., *et al.* (2007). Epigenetic stem cell signature in cancer. Nat Genet *39*, 157-158.

Wigler, M., Levy, D., and Perucho, M. (1981). The somatic replication of DNA methylation. Cell *24*, 33-40.

Wijmenga, C., van den Heuvel, L.P., Strengman, E., Luyten, J.A., van der Burgt, I.J., de Groot, R., Smeets, D.F., Draaisma, J.M., van Dongen, J.J., De Abreu, R.A., *et al.* (1998). Localization of the ICF syndrome to chromosome 20 by homozygosity mapping. American journal of human genetics *63*, 803-809.

Wolf, S.F., Jolly, D.J., Lunnen, K.D., Friedmann, T., and Migeon, B.R. (1984). Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: implications for X-chromosome inactivation. Proc Natl Acad Sci U S A *81*, 2806-2810.

Wouters-Tyrou, D., Martinage, A., Chevaillier, P., and Sautiere, P. (1998). Nuclear basic proteins in spermiogenesis. Biochimie *80*, 117-128.

Wutz, A., and Jaenisch, R. (2000). A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. Mol Cell *5*, 695-705.

Wutz, A., Smrzka, O.W., Schweifer, N., Schellander, K., Wagner, E.F., and Barlow, D.P. (1997). Imprinted expression of the Igf2r gene depends on an intronic CpG island. Nature *389*, 745-749.

Xiong, Z., and Laird, P.W. (1997). COBRA: a sensitive and quantitative DNA methylation assay. Nucleic Acids Res *25*, 2532-2534.

Xu, G.L., Bestor, T.H., Bourc'his, D., Hsieh, C.L., Tommerup, N., Bugge, M., Hulten, M., Qu, X., Russo, J.J., and Viegas-Pequignot, E. (1999). Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. Nature *402*, 187-191.

Yamada, Y., Watanabe, H., Miura, F., Soejima, H., Uchiyama, M., Iwasaka, T., Mukai, T., Sakaki, Y., and Ito, T. (2004). A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. Genome Res *14*, 247-266.

Yan, P.S., Chen, C.M., Shi, H., Rahmatpanah, F., Wei, S.H., Caldwell, C.W., and Huang, T.H. (2001). Dissecting complex epigenetic alterations in breast cancer using CpG island microarrays. Cancer Res *61*, 8375-8380.

Yan, P.S., Efferth, T., Chen, H.L., Lin, J., Rodel, F., Fuzesi, L., and Huang, T.H. (2002). Use of CpG island microarrays to identify colorectal tumors with a high degree of concurrent methylation. Methods (San Diego, Calif *27*, 162-169.

Yan, P.S., Perry, M.R., Laux, D.E., Asare, A.L., Caldwell, C.W., and Huang, T.H. (2000). CpG island arrays: an application toward deciphering epigenetic signatures of breast cancer. Clin Cancer Res *6*, 1432-1438.

Yang, A.S., Estecio, M.R., Doshi, K., Kondo, Y., Tajara, E.H., and Issa, J.P. (2004). A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements. Nucleic Acids Res *32*, e38.

Yang, P.K., and Kuroda, M.I. (2007). Noncoding RNAs and intranuclear positioning in monoallelic gene expression. Cell *128*, 777-786.

Yasui, D.H., Peddada, S., Bieda, M.C., Vallero, R.O., Hogart, A., Nagarajan, R.P., Thatcher, K.N., Farnham, P.J., and Lasalle, J.M. (2007). Integrated epigenomic analyses of neuronal MeCP2 reveal a role for long-range interaction with active genes. Proc Natl Acad Sci U S A *104*, 19416-19421.

Yen, P.H., Patel, P., Chinault, A.C., Mohandas, T., and Shapiro, L.J. (1984). Differential methylation of hypoxanthine phosphoribosyltransferase genes on active and inactive human X chromosomes. Proc Natl Acad Sci U S A *81*, 1759-1763.

Yoder, J.A., and Bestor, T.H. (1998). A candidate mammalian DNA methyltransferase related to pmt1p of fission yeast. Hum Mol Genet *7*, 279-284.

Yoder, J.A., Soman, N.S., Verdine, G.L., and Bestor, T.H. (1997). DNA (cytosine-5)-methyltransferases in mouse cells and tissues. Studies with a mechanism-based probe. J Mol Biol *270*, 385-395.

Yoon, H.G., Chan, D.W., Reynolds, A.B., Qin, J., and Wong, J. (2003). N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso. Mol Cell *12*, 723-734.

Yoshida, M., Horinouchi, S., and Beppu, T. (1995). Trichostatin A and trapoxin: novel chemical probes for the role of histone acetylation in chromatin structure and function. Bioessays *17*, 423-430.

Young, J.I., Hong, E.P., Castle, J.C., Crespo-Barreto, J., Bowman, A.B., Rose, M.F., Kang, D., Richman, R., Johnson, J.M., Berget, S., *et al.* (2005). Regulation of RNA splicing by the methylation-dependent transcriptional repressor methyl-CpG binding protein 2. Proc Natl Acad Sci U S A *102*, 17551-17558.

Yu, L., Liu, C., Vandeusen, J., Becknell, B., Dai, Z., Wu, Y.Z., Raval, A., Liu, T.H., Ding, W., Mao, C., *et al.* (2005). Global assessment of promoter methylation in a mouse model of cancer identifies ID4 as a putative tumor-suppressor gene in human leukemia. Nat Genet *37*, 265-274.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E., *et al.* (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell *126*, 1189-1201.

Zhang, Y., Ng, H.H., Erdjument-Bromage, H., Tempst, P., Bird, A., and Reinberg, D. (1999). Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. Genes Dev *13*, 1924-1935.

Zhou, Y., Cambareri, E.B., and Kinsey, J.A. (2001). DNA methylation inhibits expression and transposition of the Neurospora Tad retrotransposon. Mol Genet Genomics *265*, 748-754.

Zilberman, D. (2007). The human promoter methylome. Nat Genet *39*, 442-443.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet *39*, 61-69.

Zinn, A.R., and Ross, J.L. (1998). Turner syndrome and haploinsufficiency. Curr Opin Genet Dev *8*, 322-327.

Appendix A

Publications Arising

OPEN ACCESS Freely available online PLoS BIOLOGY

A Novel CpG Island Set Identifies Tissue-Specific Methylation at Developmental Gene Loci

Robert Illingworth[1], Alastair Kerr[1], Dina DeSousa[1], Helle Jørgensen[2], Peter Ellis[3], Jim Stalker[3], David Jackson[3], Chris Clee[3], Robert Plumb[3], Jane Rogers[3], Sean Humphray[3], Tony Cox[3], Cordelia Langford[3], Adrian Bird[1]*

1 Wellcome Trust Centre for Cell Biology, University of Edinburgh, Edinburgh, United Kingdom, **2** Lymphocyte Development Group, MRC Clinical Sciences Centre, Imperial College School of Medicine, London, United Kingdom, **3** Wellcome Trust Sanger Centre, Hinxton, Cambridge, United Kingdom

CpG islands (CGIs) are dense clusters of CpG sequences that punctuate the CpG-deficient human genome and associate with many gene promoters. As CGIs also differ from bulk chromosomal DNA by their frequent lack of cytosine methylation, we devised a CGI enrichment method based on nonmethylated CpG affinity chromatography. The resulting library was sequenced to define a novel human blood CGI set that includes many that are not detected by current algorithms. Approximately half of CGIs were associated with annotated gene transcription start sites, the remainder being intra- or intergenic. Using an array representing over 17,000 CGIs, we established that 6%–8% of CGIs are methylated in genomic DNA of human blood, brain, muscle, and spleen. Inter- and intragenic CGIs are preferentially susceptible to methylation. CGIs showing tissue-specific methylation were overrepresented at numerous genetic loci that are essential for development, including *HOX* and *PAX* family members. The findings enable a comprehensive analysis of the roles played by CGI methylation in normal and diseased human tissues.

Citation: Illingworth R, Kerr A, DeSousa D, Jørgensen H, Ellis P, et al. (2008) A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol 6(1): e22. doi:10.1371/journal.pbio.0060022

Introduction

DNA methylation in the mammalian genome arises due to covalent addition of a methyl group to the 5′ position of cytosine in the context of the palindromic dinucleotide, CpG. This modification is established and maintained by a family of DNA methyltransferases that are essential for development and viability [1,2]. The pattern of CpG methylation in the human genome distinguishes two fractions with distinct properties: a major fraction (~98%), in which CpGs are relatively infrequent (on average 1 per 100 bp) but highly methylated (approximately 80% of all CpG sites), and a minor fraction (<2%) that comprises short stretches of DNA (~1,000 bp) in which CpG is frequent (~1 per 10 bp) and methylation-free. The latter are known as CpG islands (CGIs) and they frequently colocalise with the transcription start sites (TSSs) of genes [3,4].

Although CGIs are often free of methylation, there are circumstances in which they become heavily methylated, and this invariably correlates with silencing of any promoter within the CGI. Artificial methylation of CGI promoters has long been known to extinguish transcription when the constructs are introduced into living cells [5]. Moreover, demethylation of endogenous methylated CGIs using DNA methytransferase inhibitors can restore expression of the gene [6]. These findings demonstrate that dense CpG methylation prevents expression of CGI promoters. Because of this biological consequence, it is important to know the extent of CGI methylation in both normal and diseased tissue states. The classical example is X chromosome inactivation in placental mammals, during which hundreds of CGI promoters become methylated and contribute to the stability of gene inactivation on this chromosome [7,8]. Genomic imprinting can also depend upon differential CGI methylation between maternal and paternal alleles [9]. Certain "testis-specific antigen" genes possess CGIs that are methylated in all somatic tissues, but not in testis, where the genes are expressed [10]. Several additional candidates for CGI methylation in normal tissues have been reported [11,12], and the number of cases has recently grown due to large-scale bisulfite sequencing [13] and analysis of promoter methylation using microarrays [14].

In the cases of X chromosome inactivation and genomic imprinting, the biological processes were described initially, and CpG methylation was subsequently implicated through mechanistic studies. To uncover new biological roles for CGI methylation in hitherto undiscovered biological processes, it would be advantageous to comprehensively screen genomic DNA for methylated CGIs in normal or diseased cell types. A persistent limitation affecting this kind of approach has been uncertainty concerning CGI identification [15]. The criteria for designating a sequence as CGI-like are currently exclusively bioinformatic in nature, relying on the differences in the base composition and CpG frequencies (observed/expected) between bulk genomic DNA and CGIs [16,17]. In an

Academic Editor: Edison T. Liu, Genome Institute of Singapore, Singapore

Received October 1, 2007; **Accepted** December 13, 2007; **Published** January 29, 2008

Abbreviations: CGI, CpG island; CAP, CXXC affinity purification; MAP, MBD affinity purification; TSS, transcription start site

* To whom correspondence should be addressed. E-mail: a.bird@ed.ac.uk

Author Summary

The human genome contains about 22,000 genes, each encoding one of the proteins required for human life. A particular cell type (e.g., blood, skin, etc.) expresses a specific subset of protein genes and silences the remainder. To shed light on the mechanisms that cause genes to be activated or shut down, we studied DNA sequences called "CpG islands" (CGIs). These sequences are found at over half of all human genes and can exist in either the active or silent state depending on the presence or absence of methyl groups on the DNA. We devised a method for purifying all CGIs and showed that, unexpectedly, only half occur at the beginning of genes near the promoter, the rest occurring within or between genes. Notably, methylation of CGIs causes stable gene silencing. We tested 17,000 CGIs in four human tissues and found that 6%–8% were methylated in each. Genes whose protein products play an essential role during embryonic development were preferentially methylated, suggesting that gene expression during development could be regulated by CGI methylation.

attempt to address this limitation and create a resource for future analysis, we developed a method for CGI identification and purification based on their lack of CpG methylation in an otherwise highly methylated genome.

Our method utilised a protein domain with a specific affinity for clustered nonmethylated CpG sites [18,19]. Using this reagent we physically purified DNA sequences that contain clusters of nonmethylated CpG-rich DNA from human blood DNA. Large-scale sequencing of the fraction identified a CGI set that was annotated on the ENSEMBL database. We found that many CGIs in the set were not associated with promoters of annotated genes, but were either within transcription units or between genes. By arraying the intact CGI sequences, we were able to interrogate genomic DNA fractions from several human tissues in order to identify methylated CGIs. The results revealed large numbers of CGIs that are methylated in normal human tissues, many of which showed tissue-specific methylation.

Results

A Novel Technique for Purification of CpG Islands

To enrich for nonmethylated CpG-rich DNA (CpG islands), we developed the technique of CXXC affinity purification (CAP). This uses the cysteine-rich CXXC3 domain that has a high affinity for nonmethylated CpG sites [18,19]. A recombinant CXXC domain from mouse Mbd1[19] was expressed in bacteria, and its binding specificity for nonmethylated CpG sites was confirmed (Figure 1A). The CXXC domain had no detectable affinity for DNA containing only methylated CpGs or for DNA lacking CpGs altogether. We linked the CXXC domain to a sepharose matrix and confirmed that this fractionated DNA fragments according to CpG density and methylation status (unpublished data). All DNA bound to the

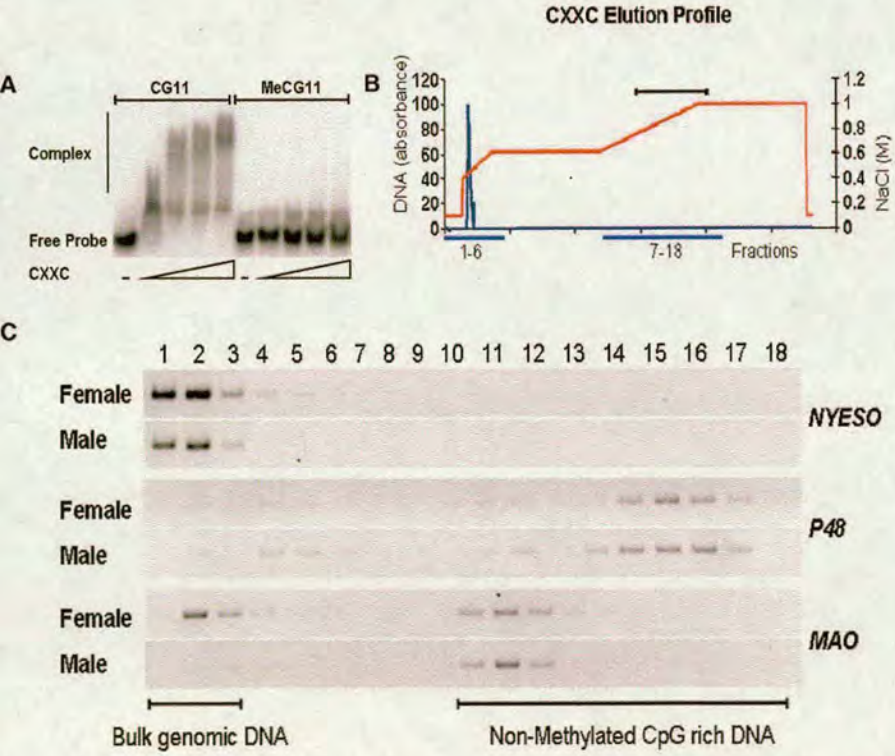


Figure 1. The Immobilised CXXC Domain Specifically Retains DNA Containing Clusters of Nonmethylated CpGs

(A) EMSA showing the CXXC complex with a DNA probe containing 27 nonmethylated CpG sites. Nonmethylated probe DNA (CG11) or methylated probe (MeCG11) was incubated with 0, 250, 500, 1,000, or 2,000 ng of recombinant CXXC protein.

(B) A typical elution profile of bulk genomic DNA (blue line) from a CXXC affinity chromatography column. Genomic DNA (100 μg) was applied to the CXXC affinity matrix (see Methods) in low salt (0.1 M NaCl) and eluted with a gradient of increasing NaCl (red line; see text). Eighteen fractions were interrogated by PCR (blue lines). The bracket above indicates fractions that were found to contain nonmethylated CGIs.

(C) Elution of specific CGI sequences of known methylation status. Methylated CGIs (*NYESO* and *MAO* in females) coelute with bulk genomic DNA (see bracket) whereas nonmethylated CGIs (*P48* and *MAO*) elute at high NaCl concentration.

doi:10.1371/journal.pbio.0060022.g001

column at 0.1 M salt. Methylated DNA and CpG-poor DNA eluted at ~0.4 M NaCl, whereas elution of nonmethylated CpG-rich DNA required 0.6–1.0 M NaCl. To test the behaviour of CGIs on the column, human genomic DNA was digested with MseI (TTAA) [20] and fractionated over the CXXC column (Figure 1B). The reasoning behind use of MseI [20] was to cut AT-rich bulk genomic DNA into small fragments (predicted average = 123 bp), but to leave CGIs relatively intact (predicted average = 625 bp). As bulk genomic DNA has a CpG on average every 100 bp, most MseI fragments will have too few CpGs to be retained by the CXXC matrix. CGIs on the other hand, with 1 CpG per ~10 bp, will give rise to long fragments with many CpGs. Eluted fractions were interrogated by PCR using primers specific for a range of known CGIs and non-CGI sequences (Figure 1C). For example, the nonmethylated CGI of the *P48* gene eluted at high salt. The X-linked monoamine oxidase (*MAO*) gene eluted as a single high salt peak from male genomic DNA (where it is nonmethylated), but as two separate peaks at high and low salt when female DNA (with one methylated and one nonmethylated allele) was fractionated. The CGI associated with the *NYESO* testis-specific antigen gene (methylated in somatic tissues) eluted from the CXXC column by low salt as predicted. The data confirm that CAP may be used to purify a CGI fraction from human genomic DNA.

A Comprehensive CGI Set from Human Blood

Most or all CGIs are in a nonmethylated state in sperm, but in addition repetitive elements [21] and telomere-proximal sequences [22], both of which are moderately CpG-rich, are hypomethylated in sperm DNA. To avoid contamination of the CGI fraction with sequences that are nonmethylated, specifically in germ cells, whole human blood was used as a source of CGI fragments. Pooled whole blood DNA from three males was fractionated using the CXXC column. High salt fractions were pooled, diluted, and re-chromatographed before cloning in plasmids. The resulting blood CGI library was analysed by 221,860 sequence reads representing 119,487 genomic templates. These compiled to give 28,013 unique MseI fragments. Plots of DNA insert length versus either G+C content or observed/expected CpG frequency (CpG[o/e]) showed that the great majority of clones exhibited a higher G+C content (average = 62%) and CpG[o/e] (average = 0.71) than bulk genomic DNA (G+C = 41% and CpG[o/e] = 0.2) (Figure 2A and 2B). A fraction of small fragments with sequence characteristics resembling bulk genomic DNA was detected by these plots. As these probably represent contamination, we filtered out fragments shorter than 512 bp that had a GC content less than 50% and/or a CpG[o/e] less than 0.6 (see grey dots in Figure 2A and 2B). The resulting final sequenced set corresponds to 17,387 CGIs and is annotated on the ENSEMBL genome browser (http://www.ensembl.org/index.html. DAS sources: "CPG island clones"). The great majority have classical CGI properties (Figure 2C). Due to their high average GC content, the sequence pass rate was 69%. Assuming that the unsequenced clones reflect the same proportion of CGIs as those that were sequenced, we estimate the total number of CGIs in the library as 25,200. It is likely that a higher proportion of sequence failures affect bona fide CGIs, as GC-richness is known to interfere with sequencing. If so, we estimate that the number of human genomic CGIs may be closer to 30,000.

CGIs are identified bioinformatically as DNA sequences with a base composition greater than 50% G+C and a CpG[o/e] of more than 0.6 [23]. The DNA length over which this condition applies is critical. Initially the threshold most often used was 200 bp, whereas 500 bp is now more commonly applied [17]. These two criteria are formalised as "NCBI-relaxed" and "NCBI-strict," respectively (http://www.ncbi.nlm.nih.gov/mapview/static/humansearch.html#cpg). The relaxed algorithm predicts 307,193 CGIs in the human genome, which includes many repeated sequences and gene exons. Over 90% of NCBI-relaxed CGIs are not represented in either our library or the set predicted by the NCBI-strict. This and other arguments suggest that the great majority (>90%) are false positives. On the other hand, 77% of clones in the CGI library match CGIs predicted by the "NCBI-strict" algorithm (Table 1). Examples of the coincidence of NCBI-strict predicted CGIs and sequenced CGI clones are illustrated for the three typical regions of the human genome (Figure 2D).

Altogether, NCBI-strict identifies 24,163 CGIs in the human genome, which accords with the adjusted CGI library estimate of 25,200. The coincidence of these numbers masks significant differences, however, as 23% of CGIs in the library are not detected by the NCBI-strict algorithm (4,082 out of 17,387; Table 1). Four randomly selected examples of library CGIs not detected by NCBI-strict (Figure 2D and 2E, numbered) gave CpG maps resembling CGIs; three of these coincided with the promoters of annotated protein-coding genes (Figure 2D and 2E). The presence of bioinformatically predicted CGIs that are missing from the CGI library is most probably due to sequence failure of ~31% of library inserts. Analysis of the CGIs missed by the NCBI-strict algorithm shows them to be, as expected, significantly weaker with respect to CpG and G+C content than the total set (Figure S1). It was not obvious, however, that the algorithm could be easily improved based on this information. Relaxation of the sequence parameters reduces the number of false negatives, but leads to increased numbers of false positives. We suggest that CAP identifies islands that fail the NCBI criteria, but reduces the false discovery rate by excluding spurious methylated CpG-rich sequences. Like the majority of CGIs, most NCBI-missed islands are gene-associated, although with an increased incidence of intragenic islands (Table S1). The CGI library therefore includes a significant fraction of bona fide CGIs that are missed by one of the best available algorithms.

CAP defines a set of CGIs that is coherent with respect to clustering of nonmethylated CpG sites. The genomic distribution of these CGI sequences correlates strongly with gene density (Figure 2F). For example, gene-rich Chromosome 19 is also CGI-rich, whereas gene-poor Chromosome 18 is correspondingly CGI-poor. With respect to annotated protein-coding genes, we found that 76% of CGIs are within 1.5 kb of a transcription unit, but only 49% overlap with the TSS (Table 2). It follows that half of CGIs are not TSS-associated, but are either within downstream regions of transcription units (22%) or located in intergenic DNA. Previous studies have detected CGIs at the TSS of 56% of human protein-coding genes [24]. As 43.5% of TSSs overlap sequenced CGIs, we calculate that the sequenced set of 17,387 CGIs represents 78% of the CGI complement. According to this calculation, the total CGI number would be 22,400, somewhat less than

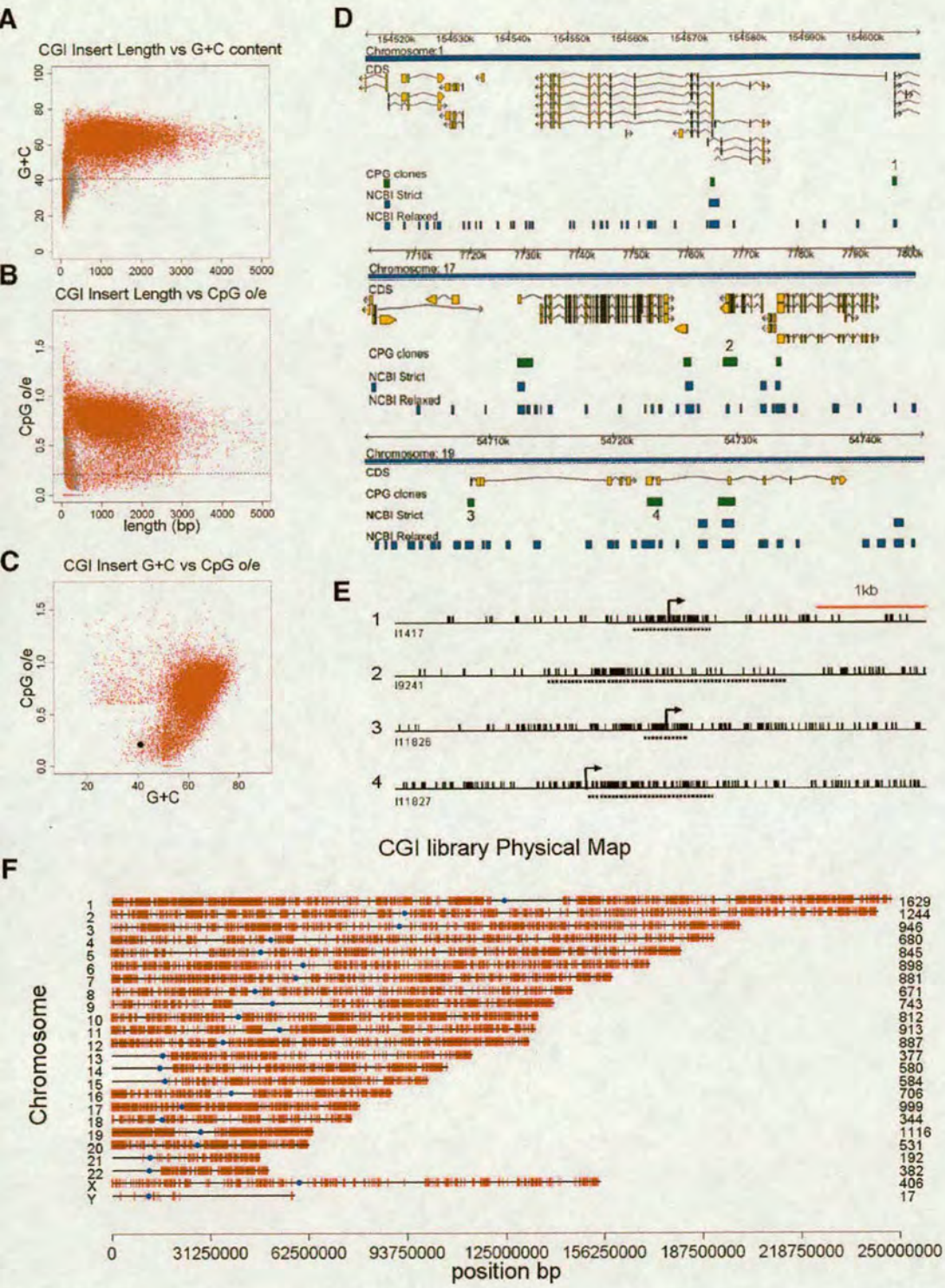


Figure 2. A Library of DNA Sequences that Bind Tightly to the CXXC Column Represents a Comprehensive Set of CGIs

(A and B) Plots of fragment length versus G+C content (A) and CpG[o/e] (B) for 28,013 unique Mse1 inserts. Fragments shorter than 512 bp with a G+C content = <50% and a CpG[o/e] = <0.6 (grey dots) were filtered out as contamination. The dashed line indicates the base composition (A) and CpG o/e (B) of bulk genomic DNA.

(C) A filtered insert set representing 17,387 CGIs shows a discrete distribution that is distant from bulk genomic DNA (black dot).

(D) Three random chromosomal regions showing CGI sequences mapped by ENSEMBL (green bars). Also shown are CGIs predicted by the NCBI-strict and NCBI-relaxed algorithms (blue bars). The directions of transcription of coding sequences (yellow bars) are arrowed. Numbered CGIs (1–4) represent sequences not detected by the NCBI-strict algorithm.

(E) CpG maps of the four CGI clones not predicted by NCBI-strict. Transcription start sites in examples 1, 3, and 4 are indicated by arrows. Sequenced MseI fragments are denoted by dashed lines and CpG sites by vertical black strokes.

(F) The distribution of cloned CGIs (red strokes) on human chromosomes. The number of CGIs on each chromosome is shown (right) and centromeres are denoted by blue dots.

doi:10.1371/journal.pbio.0060022.g002

Table 1. Comparison of Human Blood CGI Set with Bioinformatic Prediction

Criterion	Number of CGIs	$NCBI^{relaxed}$	$NCBI^{strict}$	CGI
$NCBI^{relaxed}$	307,193	307,193	24,163 (7.9%)	16,072 (5.2%)
$NCBI^{strict}$	24,163	24,163 (100%)	24163	13,568 (56.2%)
CGI	17,387	15,799 (90%)	13,305 (76.5%)	17,387

http://www.ncbi.nlm.nih.gov/mapview/static/humansearch.html#cpg
doi:10.1371/journal.pbio.0060022.t001

the figure of 25,200 deduced from the fraction of sequenced inserts.

MBD Affinity Purification and Blood CGI Methylation

CAP selects CGIs from blood DNA based on their lack of methylation and therefore excludes the small fraction of CGIs (<3%) that are fully methylated in somatic cells from the set [14]. Indeed, CGIs associated with the human testis-specific antigen genes [10], which are methylated in somatic tissues, were not enriched by CAP (Figure 1C) or present in the library (unpublished data). Despite the absence of these fully methylated CGIs, we reasoned that the blood CGI library provides an opportunity to screen for methylation that affects a fraction of all copies of a specific CGI in whole blood DNA. Also, it permits a screen for differential methylation of CGIs in tissues and cell types other than blood. To investigate CGI methylation in normal human tissues, we constructed an array of sequenced CGIs from the library by immobilising single-stranded PCR-amplified inserts on glass slides using 5′-aminolink chemistry as described (http://www.sanger.ac.uk/Projects/Microarrays/arraylab/methods.shtml). As probes for the array, methylated CGIs were enriched from genomic DNA using MBD affinity purification (MAP), which was shown previously to efficiently bind methylated CGIs [20] (Figure 3A and 3B). Human male and female blood DNA was MseI-digested and ligated to universal catch linkers. We verified by PCR that affinity fractionation using MAP effectively separated known methylated CGIs (*XIST* on the active X chromosome and *NYESO*) from bulk genomic DNA and nonmethylated CGIs (*P48* and *XIST* on the inactive X chromosome; see Figure 3B). Male and female DNA fractions were pooled after two rounds of MAP, amplified by linker-mediated PCR, cyanine labeled, and hybridized to the CGI microarray. Quadruplicate hybridisations (inclusive of cyanine dye swaps) gave mean enrichment values (MAP/Input) that allowed a comparison between male and female methylated CGI complements. As expected, these were positively correlated ($R = 0.865$ Pearson correlation) suggesting similar overall patterns. As the library comprises MseI fragments that sometimes overlap minimally with the cognate CpG-rich region, we chose to disregard data from spots that contained DNA with an average CpG frequency (observed/expected) of less than 0.5. Although the omitted fragments often denote CGIs, they include too little of the CpG-rich domain to be reliable for detection of MAP probes. This refinement reduced the number of analysable CGIs on the array to 14,318. To assess the relationship between hybridization signal relative to input and degree of enrichment by MAP, we measured a selection of CGIs in the probe by quantitative PCR and compared this data with the M values (log2 [MAP signal]/[Input signal]) for those sequences (Figure 3C). The results established that M values greater than 1.5 denote CGIs that are significantly enriched by MAP and therefore methylated. CGIs of the *BEST1* and *R4RL1* genes were predicted to be nonmethylated (M = 0.2–0.4) and methylated (M = 2.2–2.8), respectively, based on the array data. Bisulfite genomic sequencing confirmed this expectation (Figure 3G and 3H).

The major difference in CGI methylation between male and female DNA was expected to be due to X chromosome inactivation (see also [25]). We therefore compared the methylation status of CGIs on Chr 16 and Chr X in male versus female DNA. Chr 16 CGIs did not vary between males and females, whereas Chr X CGIs were significantly enriched in female DNA as predicted (Figure 3D–3F; Table S2). Studies of human X chromosome inactivation have indicated that a proportion of genes escape inactivation and are therefore expressed from both chromosomes [26,27]. By comparing the microarray data for a set of inactivated and escaping CGIs, we found that inactivated genes had significantly higher M values (p-value = 1.213×10^{-5}) (Figure 3I). This finding affirms the long-standing link between CGI methylation and gene silencing and validates the present experimental system as a means of detecting genes that are shut down in this way.

Table 2. Relationship between CGI Library Inserts and Protein-Coding Genes

Type of Overlap	Gene Type	Total Genes (CGIs)	Overlap with CGI (Gene)	Percentage	Overlap with CGI (TSS)	Percentage
Gene Overlap: CGI	Protein	21,384	15,118	70.7	9,312	43.5[a]
	All genes[b]	31,524	15,433	49.0	9,529	30.2
CGI Overlap: Genes	Protein	**17,387**	**13,271**	76.3	**8,491**	48.8
	All genes[b]	**17,387**	**13,360**	76.8	**8,611**	49.5

[a]This fraction is less than the known fraction of genes with promoter CGIs (56%), because 31% of CGI inserts did not yield DNA sequence.
[b]All genes as classified on the ENSMBL genome browser including noncoding RNAs, pseudogenes, VDJ regions, etc.
doi:10.1371/journal.pbio.0060022.t002

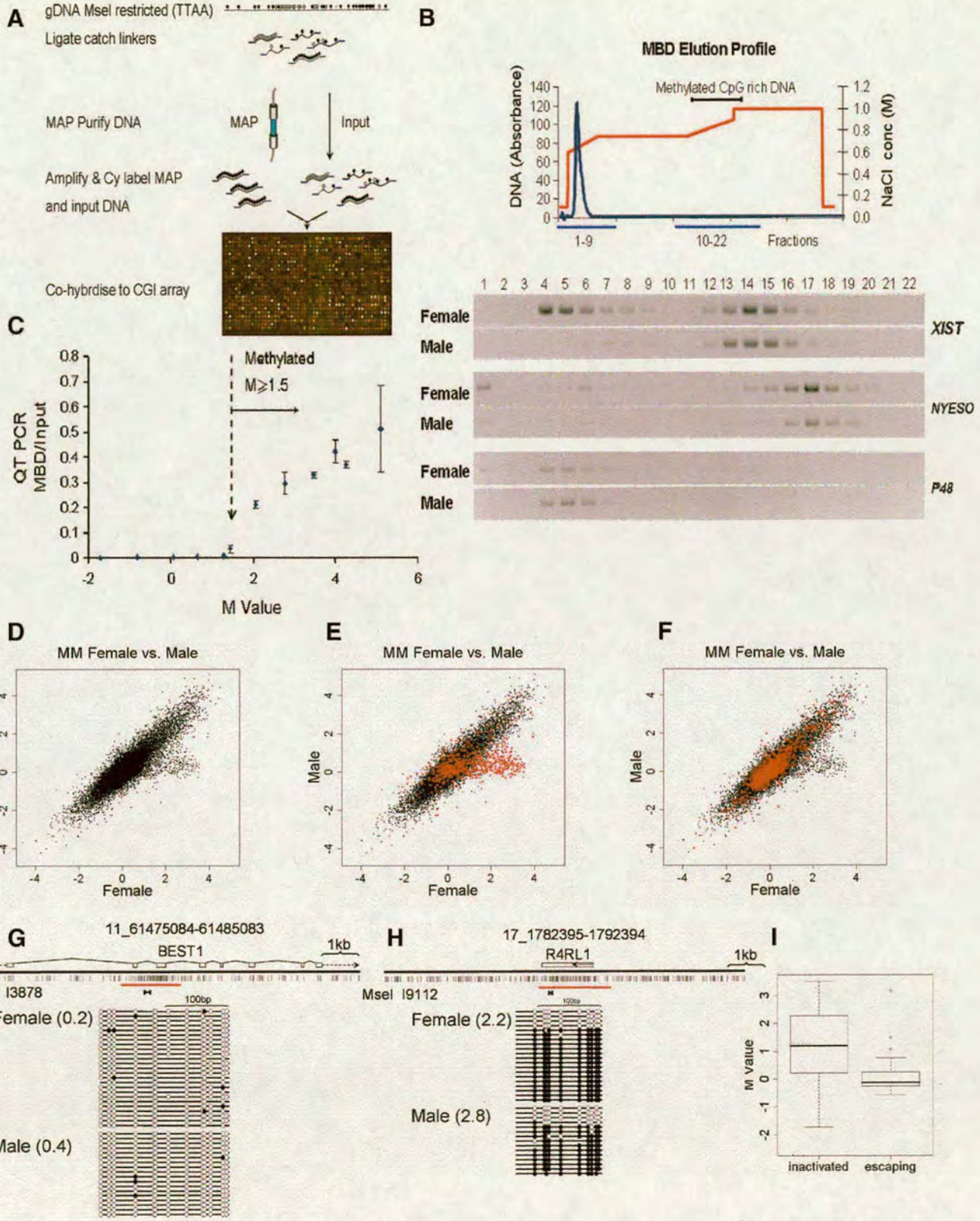


Figure 3. Use of an Arrayed CGI Library to Detect CGI Methylation in Human Blood DNA

(A) Schematic showing isolation of densely methylated CGIs using MBD affinity purification based on reference [20]. Open and filled circles represent nonmethylated and methylated CpG sites, respectively.

(B) Examples of retention of known methylated CGIs by MBD affinity chromatography. Methylated *XIST* and *NYESO* CGIs elute at high salt concentration, whereas nonmethylated *P48* and female *XIST* co-elute with bulk genomic DNA (blue line) at low salt concentration (red line).

(C) *M* values ($\log_2$[MBD/Input]) >1.5 (dashed vertical arrow) denote DNA fragments enriched by MAP. *M* values are plotted against the ratio of fragment abundance in the MAP probe versus input DNA as determined by quantitative PCR. Error bars represent ± standard deviation.

(D–F) MAP CGI array hybridization identifies CGIs that are methylated on the inactive X chromosome. (D) Probes isolated by MAP from male and female

whole blood DNA detected female-specific CGI methylation. (E) CGIs on the X chromosome (red dots) often showed female-specific methylation. (F) CGIs on Chromosome 16 (red dots) were indistinguishably methylated between male and female.
(G and H) Confirmation of methylated CGIs by bisulfite genomic sequencing. CGI clones I1387 (G) and I9112 (H) are nonmethylated and methylated, respectively, as predicted by the microarray data. Open and filled circles represent nonmethylated and methylated CpG sites, respectively. The genomic locus including annotated transcripts and CpG maps (vertical strokes) are shown above each profile. Each column represents products of amplification by a single primer pair (brackets below CpG map). Each line corresponds to a sequenced DNA strand. Red bars indicate the location of the MseI fragment cloned in the CGI library.
(I) The CGI array distinguishes genes inactivated on the X chromosome (inactive) from genes that escape inactivation (escaping). CGIs associated with inactivated genes ($n = 103$) show significantly higher M values than CGIs at escaping genes ($n = 14$; KS test: $p = 1.2 \times 10^{-5}$).
doi:10.1371/journal.pbio.0060022.g003

Differential CGI Methylation in Human Tissues

Methylation of CGIs on the inactive X chromosome and at imprinted genes is well known, but CGI methylation at other chromosomal loci in normal cells and tissues is incompletely characterized [12,13,28,29,30]. To investigate this issue on a large scale, we probed CGI arrays with MAP fractions from genomic DNA (three individuals per pool) of brain, muscle, spleen, and sperm in addition to blood (Figure 4A). MAP enrichment of methylated CGIs in sperm DNA consistently failed to generate enough DNA for labeling using our standard PCR amplification conditions and was therefore not analysed further. We conclude that the level of CGI methylation in sperm is far lower than in any of the somatic tissues. Taking M values greater than 1.5 to signify methylation, we observed between 5.7% and 8.3% of CGIs methylated in the somatic tissues that were tested (Figure 4B; Table 3; Dataset S1). Some CGIs were methylated in common between all the tested somatic tissues, whereas others were methylated in only one or a subset of the tissues. We noted that methylated CGIs disproportionately involved those that are remote from the TSS of an annotated gene. In the dataset as a whole, only 8% of TSS CGIs showed evidence of methylation in at least one tissue, whereas 22% of 3′ CGIs were methylated (Table 4). Do the methylated CGIs differ in sequence characteristics from CGIs that remain methylation-free? We plotted the CpG[o/e] frequencies of 1,657 CGIs that acquired methylation in one or more tissues and found a mean CpG[o/e] of 0.77 compared with 0.75 for methylated CGIs (Figure 4C). Though statistically significant (p-value = 1.413e-10) the biological significance of this small difference is unclear.

We checked by bisulfite sequence analysis a panel of seven CGIs with M values suggestive of tissue-specific methylation (M values differing between tissues by >0.75). In each case, bisulfite data confirmed the microarray predictions. CGI I1878 is not associated with an annotated gene (±1.5 kb) and is methylated exclusively in muscle and brain (Figure 4D). CGI I2985 spans the transcription start site of the *SEC31B* gene, whose product is implicated in vesicular trafficking, and is compositely methylated only in blood and spleen (Figure 4E). CGIs I13406 (Figure 4F) and I12175 (Figure 5A) are methylated specifically in muscle. These overlap the predicted gene 67313 and the 3′ end of *OSR1*. CGI I3654, which is associated with the promoter region of an annotated *PAX6* transcript (Q59GD2), previously shown to contain methylated CpG sites [31], is specifically methylated in brain (Figure 5B). I11878 is a 3′ CGI of *ZN649* and is only methylated in spleen (Figure 4G).

Many methylated CGIs were associated with genes that are essential for development (Figure 5). This was confirmed by analysis of gene ontology, which showed significant over-representation of genes whose products are involved in developmental processes, including ectoderm and mesoderm development, neurogenesis, and segment specification (Table S3). Transcription factors, including homeobox family members and other DNA binding proteins, were twice as abundant as expected by chance. Other gene categories did not show significant enrichment. Among the CGIs whose methylation status was confirmed by bisulfite sequencing, *PAX6* is involved in eye development and neurogenesis [32], the *HOXC* cluster lays down the embryonic body plan, and *OSR1* is related to a gene involved in *Drosophila* gut development. We examined the extended *HOXC* and *PAX6* loci for CGI methylation status using the MAP-CGI array data. Our library identified 19 CGIs within the 150-kb *HOXC* gene cluster of which eight were methylated differentially in blood, muscle, and spleen (Figure 5C). Brain was the only tissue that lacked obvious *HOXC* CGI methylation. Of nine CGIs near *PAX6*, two showed differential methylation. In addition to brain-specific methylation of the *PAX6-Q59GD2* CGI (see Figure 5B), we observed methylation of a CGI upstream of the major *PAX6* promoter in muscle and brain (Figure 5D).

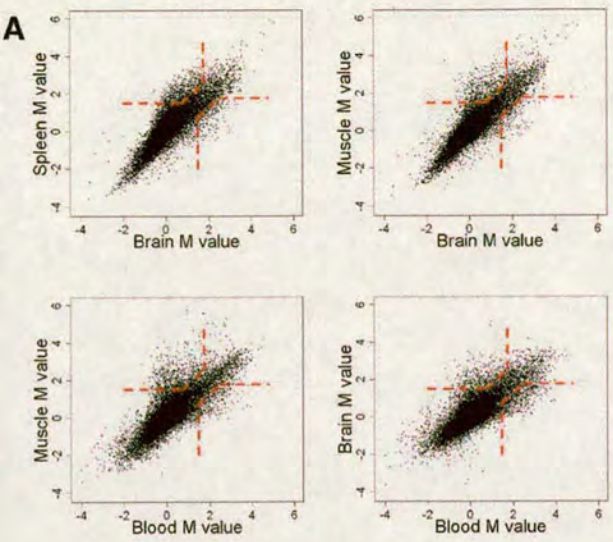
The majority of CGIs identified as methylated by MAP-CGI array hybridization display composite methylation (Figures 3, 4, and 5), whereby DNA strands at a specific locus were either heavily methylated or essentially nonmethylated. This can explain why CGIs that were initially selected by being nonmethylated in blood DNA (by CAP) nevertheless register as methylated by MAP-CGI array analysis. One potential explanation for composite CGI methylation is that different individuals within the tissue pools exhibit different CGI methylation. To look for such "polymorphism," we examined CGI I5134, which is within the *HOXC* cluster and shows composite methylation by bisulfite genomic sequencing. Analysis of individuals by MAP-CGI arrays showed highly significant differences between individual C and individuals A and B (Figure 5E). This strikingly confirms individual variability in methylation at this CGI.

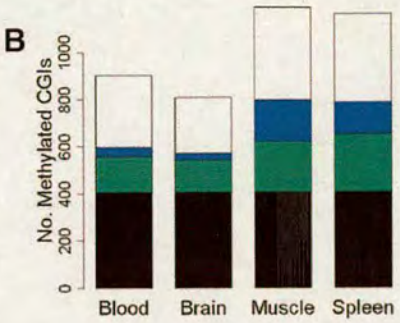
Another potential explanation for composite CGI methylation is that cell types within the tissue sample possess different CGI methylation profiles. Blood, for example, consists of monocytes and granulocytes, each of which is subdivided into other cell types. As CGI I2985 was methylated at about half of DNA strands in blood, we tested the level of CGI methylation in DNA from monocytes and granulocytes separately. The results showed that monocytes had high methylation levels at this CGI, whereas granulocytes had very low methylation (Figure 5F). These findings indicate a developmental origin for cell type–specific methylation at this genomic CGI.

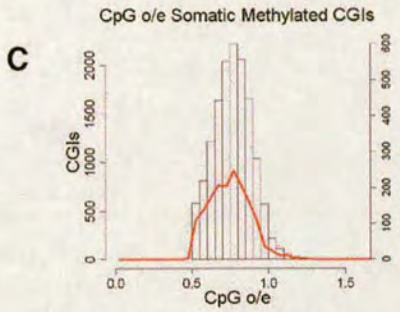
Discussion

A Comprehensive CGI Set

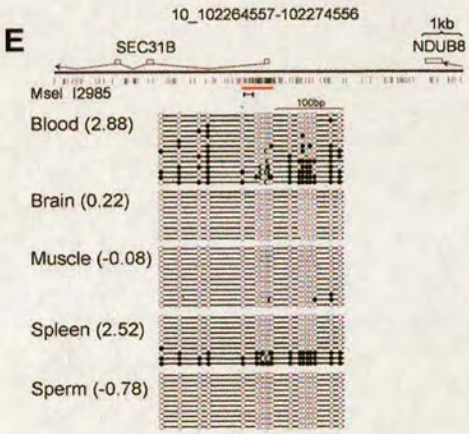
We describe the characterisation of a comprehensive, verified CGI set derived from human blood genomic DNA

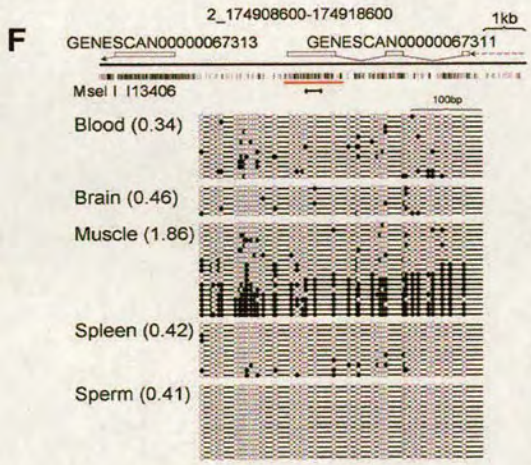

A
Spleen M value
Brain M value
Muscle M value
Brain M value
Muscle M value
Blood M value
Brain M value
Blood M value


B
No. Methylated CGIs
Blood
Brain
Muscle
Spleen


C
CpG o/e Somatic Methylated CGIs
CGIs
CpG o/e


D
1_208487900-208497900
1kb
MseI I1878
100bp
Blood (-0.56)
Brain (2.20)
Muscle (2.51)
Spleen (-1.37)
Sperm (1.15)


E
10_102264557-102274556
1kb
SEC31B
NDUB8
MseI I2985
100bp
Blood (2.88)
Brain (0.22)
Muscle (-0.08)
Spleen (2.52)
Sperm (-0.78)


F
2_174908600-174918600
1kb
GENESCAN00000067313
GENESCAN00000067311
MseI I I13406
100bp
Blood (0.34)
Brain (0.46)
Muscle (1.86)
Spleen (0.42)
Sperm (0.41)

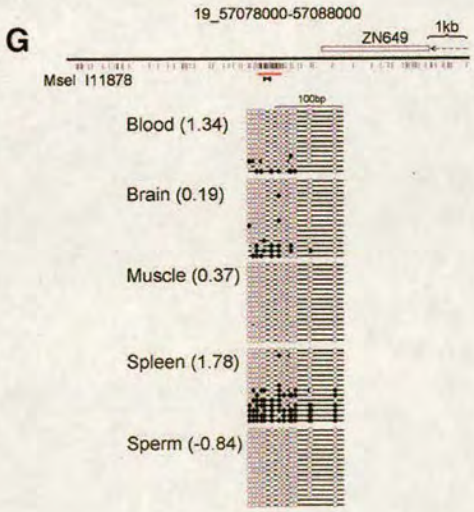

G
19_57078000-57088000
ZN649
1kb
MseI I11878
100bp
Blood (1.34)
Brain (0.19)
Muscle (0.37)
Spleen (1.78)
Sperm (-0.84)

Figure 4. Tissue-Specific CGI Methylation in a Panel of Human Tissues

(A) Examples of pairwise comparisons using MAP CGI probes derived from blood, brain, muscle, and spleen. Broken red lines indicate threshold *M* values used to determine differential CGI methylation.

(B) Frequencies of methylated CGIs in blood, brain, muscle, and spleen. The following catagories are represented: CGIs methylated in all tested tissues (black); CGIs methylated in more than one tissue tested but not all (green); CGIs methylated in one tissue only (blue); CGIs methylated in one tissue tested but unclassified in other tissues (white).

(C) Somatically methylated CGIs display a very small but significant reduction in CpG[o/e] (0.75) relative to the whole CGI set (0.77; $n = 1,657$ and 12,661, Wilcoxon rank test: *p*-value: $1.022e^{-11}$). The histogram shows the CpG[o/e] profile for the total CGI set (white bars) overlaid with the CpG[o/e] profile for methylated CGIs (red line).

(D–G) Confirmation of candidate CGIs showing evidence of tissue specific methylation by bisulfite genomic sequencing. Layout is as for Figure 3G.

doi:10.1371/journal.pbio.0060022.g004

that will be beneficial for studies of CGIs in normal human tissues and in disease settings. By focusing on CGIs alone, we excluded ~98% of the genome from our analysis. While it will ultimately be important to know in detail the methylation status of whole genomes, this currently represents a technical challenge that has been addressed only for the small-genomed plant *Arabidopsis* [33,34]. These studies used indirect microarray-based methods for mapping DNA methylation that depend upon probes enriched in methylated domains. Current enrichment methods require clusters of CpG methylation, which are notably absent from the CpG-deficient majority of the mammalian genome. As a result, much bulk genomic DNA is beyond the resolution limit of this approach. Whole genome bisulfite sequencing, the most direct and reliable method for mapping methylated sites, has not yet been attempted in any organism. We therefore decided to study a discrete genomic fraction with evident biological relevance whose methylation status can be interrogated using microarray-based methods.

To isolate nonmethylated duplex CGIs from total genomic DNA, we harnessed the binding specificity of the CXXC protein domain. Extensive sequencing of the resulting library confirmed that CGIs represent a discrete fraction of the human genome with shared DNA sequence characteristics. The present CGI set supercedes a previous human CGI library that was prepared in our laboratory using an indirect affinity purification procedure [20]. The initial library was not comprehensive and appears to have acquired significant levels of non-CGI contamination following amplifications. We estimate that the new library represents ~25,000 CGIs, of which ~60% have been arrayed as full-length single strands on glass slides. Additional analysis of inserts that initially failed conventional sequencing strategies will generate an array that covers the great majority of CGIs that are nonmethylated in human blood. The choice of blood DNA as a starting material necessarily excludes from the set any CGIs that are nonmethylated in germ cells, but densely methylated in the soma [14]. In the future, it will be instructive to compare an exhaustive sequence analysis of this set with comparable sequences isolated by CAP from sperm DNA.

Table 3. CGI Methylation in Human Tissues

Methylation Status	Blood	Brain	Muscle	Spleen
Methylated in all	408	408	408	408
Differentially methylated (multiple[a])	149	135	214	247
Differentially methylated (single[a])	50	35	178	140
Unclassified methylation	303	237	392	381
Total	910	815	1,192	1,176
CGIs	14,318	14,318	14,318	14,318
Percentage methylated	6.4	5.7	8.3	8.2

[a]Refers to number of tissues tested.

doi:10.1371/journal.pbio.0060022.t003

The library prepared using CAP defines CGIs based on the empirical criterion of clustered nonmethylated CpGs, whereas criteria based purely on base sequence and composition necessarily ignore methylation status. Comparing our set with predicted CGIs on the NCBI database shows good overlap with predictions based on the "strict" algorithm. The CGI library did, however, identify 23% of CGIs that were negative by this criterion. This suggests that the software for DNA sequence-based CGI identification misses almost one in four CGIs that the more biological criterion of CAP is able to include. Recent CGI analyses identified large numbers of human CGI promoters that are enriched in methylation at lysine 4 of histone H3, a mark of transcriptional activity [14,35,36]. Since it has been proposed that hypomethylation is dependent on germ line and early embryonic transcription [3], we determined the overlap between our CGI set and the H3K4 sites in human embryonic stem cells [37]. We calculate that 90% of CGIs in the filtered set (14,318) coincide with H3K4 methylated promoters that were reported in the chromatin study. A better test of the relationship between CGIs and H3K4 methylation islands in ES cells is to exclude promoters of annotated genes and focus on intra- and intergenic CGIs. Here again, a high proportion (75%) of CGIs overlap with H3K4 methylation islands. These findings are compatible with the notion that the presence of CGIs is connected with specialised chromatin configurations in early embryonic cells. An intriguing proposal is that H3K4 methylation may be incompatible with docking of de novo methyltransferases [38]. This could in theory insure that these regions remain free of CpG methylation at a time when the rest of the embryonic genome is subject to global methylation.

We found that 49% of CGIs overlap the TSS of an annotated gene. In considering the function of the half of CGIs that are remote from an annotated TSS, it is noteworthy that several intragenic CGIs have been shown to coincide with previously unforeseen promoters that initiate bona fide transcripts [39,40]. This raises the possibility that all CGIs function as promoters and are therefore TSS-associated [40]. In this connection, it is of interest that genome-wide analysis by tiling arrays detected over 10,000 unanticipated human transcripts, many of which may represent noncoding RNAs [41]. It is conceivable that many inter- and intragenic CGIs mark promoters that drive the synthesis of these novel transcripts. The noncoding transcripts *XIST* and *AIR*, for example, whose RNA products play regulatory roles [42–44], both initiate within CGI promoters. The proximity of many methylated CGIs to developmentally important genes raises

Table 4. Location of Methylated CGI Relative to Protein-Coding Genes

CGI Gene Association	All CGIs	Methylated	Methylated (%)	Differentially Methylated	Differentially Methylated (%)
All CGIs	14,318	1,657	11.6	711	5
Genes	11,383	1,209	10.6	501	4.4
5′	7,915	620	7.8	243	3.1
3′	765	166	21.7	77	10.1
Intragenic	3,478	536	15.4	230	6.6
Intergenic	2,863	435	15.2	203	7.1

doi:10.1371/journal.pbio.0060022.t004

the possibility that putative CGI transcripts play regulatory roles during development. Recent analyses of the human *HOX* gene cluster highlight the functional importance of noncoding RNAs [45]. Large numbers of potential CGI promoters within *HOX* gene loci may therefore contribute to the regulation of these complex loci.

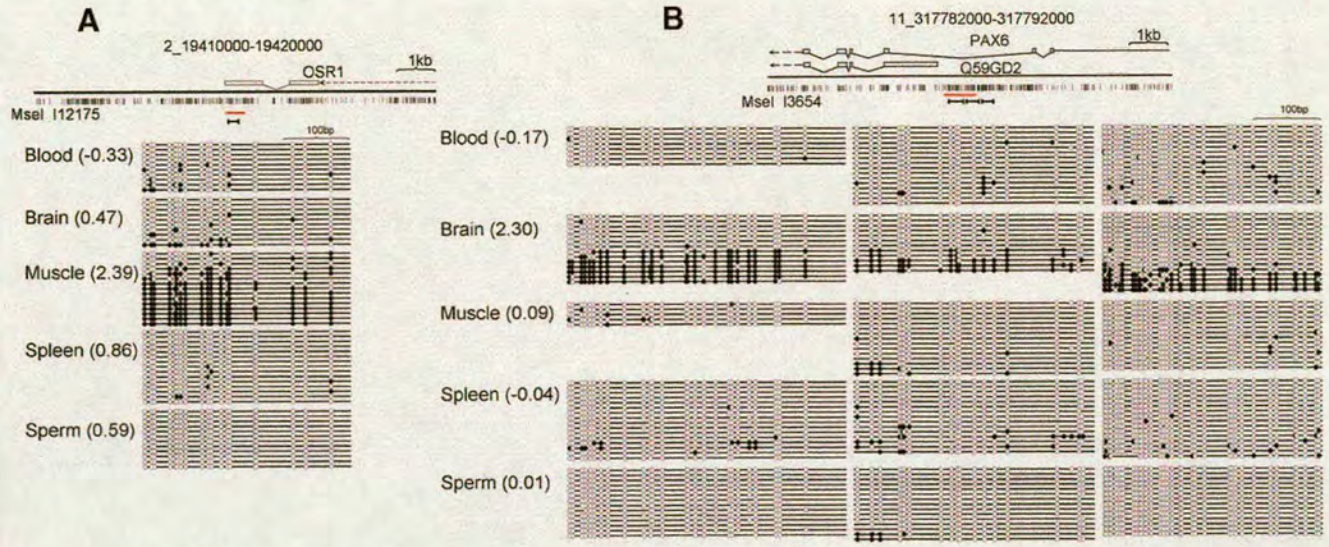
CGI Methylation in Normal Tissues

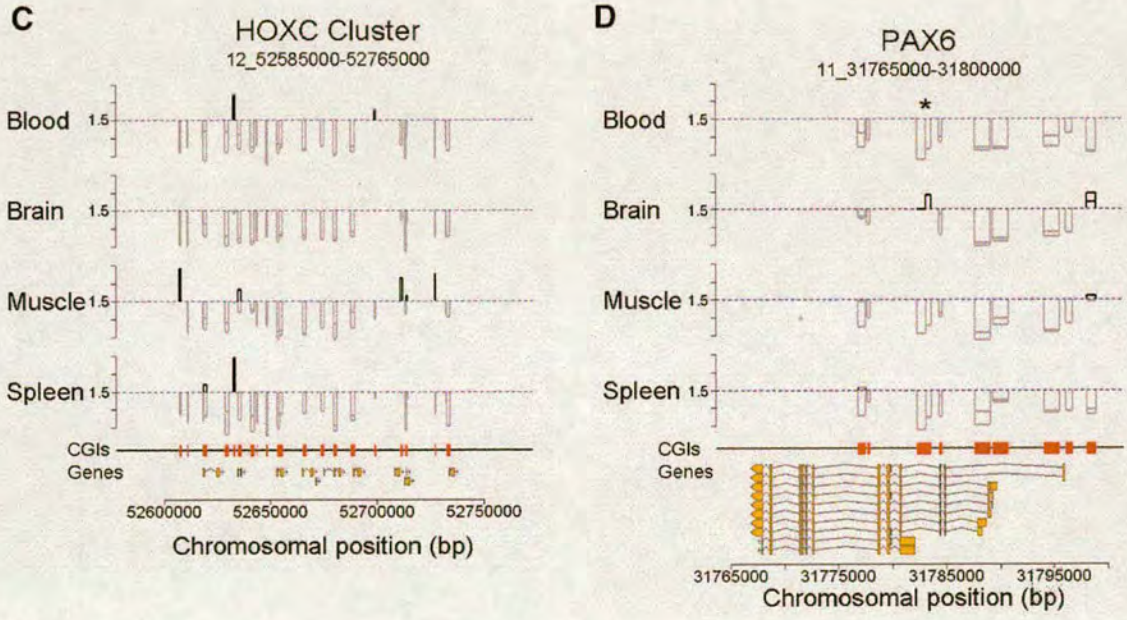
CGI methylation has been extensively studied in cancers and their derivative cell lines, but relatively less attention has been paid to the phenomenon in normal tissues. Several studies have reported somatic CGI methylation, but in early examples the bioinformatics procedure used to identify these sequences was often equivalent to the NCBI-relaxed algorithm, which generates a large excess of questionable CGI candidates. The *MASPIN* gene, for example, scores as a methylated CGI promoter by the relaxed criterion [28], but it is not detected as such either by the NCBI-strict algorithm or by CAP (unpublished data). A recent report addressing the methylation status of 16,000 human promoters identified that 3% of TSS-associated CGIs are normally methylated in somatic tissue [14], which is somewhat below the levels observed in our study (7.8%; Table 4). We detect a much higher frequency of methylation at nonpromoter CGIs (average = 16%), which are obviously absent from promoter arrays. In particular, 22% of CGIs near the 3′ ends of genes are methylated. Extensive bisulfite sequence analysis [13] surveyed 512 CGIs on Chrs 6, 20, and 21 and reported 9.2% to be methylated in somatic tissues. This is similar to the overall level of 11.6% methylation among 14,318 CGIs detected by our study (Table 4).

Our findings raise important questions about the relationship of CGI methylation to gene expression. On the X chromosome, it is clear that methylated CGIs correlate with inactivated genes whereas unmethylated CGIs correlate with genes known to escape inactivation. The generalisation that CGI methylation silences promoters is therefore supported (see also [25]). The relevance to gene expression of the autosomal methylated CGIs identified here is complicated by the frequent presence of both methylated and nonmethylated alleles in a specific tissue (see below). This means that even if CGI methylation silences a promoter completely, large changes in gene expression are not to be expected. Also, many CGIs are not at promoters of annotated genes, but are within or between transcription units. Their function with respect to transcription, if any, may be positive or negative. Finally, any transcripts originating from these "orphan" CGIs have yet to be identified and cannot be tested. For these reasons, it is difficult to make predictions about the effect of CGI methylation on global transcription levels. We nevertheless mined published expression microarray data to determine whether tissues in which a specific set of promoter CGIs was methylated expressed the associated genes at a different level from tissues where the same CGI was unmethylated. The results showed no obvious correlation between CGI methylation and expression. This, therefore, remains an open question that demands detailed analysis of specific cases.

Genes that play an important role in development were prominent among the set of methylated CGIs identified by MAP-CGI array hybridization. Out of 109 CGI-associated genes that contain homeobox-like domains, 27 (~25%) were unmethylated in at least one tissue compared with ~11% of all CGI-associated genes (see Table 4). Specifically, we identified 79 CGIs in the four human *HOX* gene clusters A–D, of which 22 were methylated in at least one of the tissues that we tested. Given the relatively small selection of tissues analysed in the study, the actual frequency of *HOX* CGI methylation in all human tissues is likely to be higher than one in four. Interestingly, methylation of *HOX* gene CGIs is also reported in cancers [46], raising the possibility that cancer CGI methylation patterns mimic patterns that arise during development. A potential link between normal development and cancer is suggested by the finding that CGIs methylated in cancer preferentially include promoters that are marked by association with polycomb group proteins in embryonic stem cells [47–49]. In contrast, we found little difference between the fractions of all CGIs (5.9% = 845/14,318) and of methylated CGIs (7.7% = 127/1,657) that were polycomb-associated in embryonic cells [37]. The origins of CGI methylation in cancer may be distinct from the mechanisms that lead to CGI methylation in normal tissues.

It was reported that the most CpG-rich CGIs among 512 analysed on Chr 6, Chr 20, and Chr 22 were never methylated, suggesting that the CpG-richness may protect from methylation [13]. In a larger CGI set, we detected a very small, but statistically significant, difference in sequence properties between CGIs that become methylated and those that remain immune in the tested cell types. The mean CpG[o/e] was 0.75 for methylated CGIs compared with 0.77 for bulk CGIs (Figure 4C). Bock and colleagues [50] identified sequence features that were predictive for CGI methylation, including specific repeats, sequence patterns, and DNA structure. Contrary to predictions of this method, methylated CGIs were significantly depleted in repetitive elements and showed no difference in predicted base twist. We did, however, observe small, but statistically significant, increases in simple


A
2_19410000-19420000
OSR1
1kb
MseI I12175
100bp
Blood (-0.33)
Brain (0.47)
Muscle (2.39)
Spleen (0.86)
Sperm (0.59)
B
11_317782000-317792000
PAX6
1kb
Q59GD2
MseI I3654
100bp
Blood (-0.17)
Brain (2.30)
Muscle (0.09)
Spleen (-0.04)
Sperm (0.01)


C
HOXC Cluster
12_52585000-52765000
Blood
Brain
Muscle
Spleen
1.5
CGIs
Genes
52600000
52650000
52700000
52750000
Chromosomal position (bp)
D
PAX6
11_31765000-31800000
Blood
Brain
Muscle
Spleen
1.5
CGIs
Genes
31765000
31775000
31785000
31795000
Chromosomal position (bp)

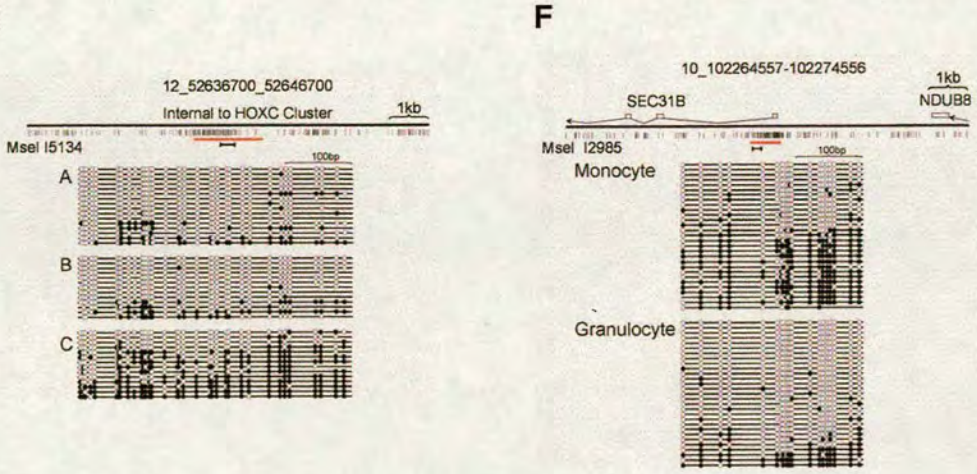

E
12_52636700_52646700
Internal to HOXC Cluster
1kb
MseI I5134
100bp
A
B
C
F
10_102264557-102274556
SEC31B
1kb
NDUB8
MseI I2985
100bp
Monocyte
Granulocyte

Figure 5. Tissue, Cell-Type, and Individual-Specific CGI Methylation at Developmental Gene Loci

(A–B and E–F) Bisulfite genomic sequencing confirmed tissue-specific CGI methylation associated with the developmental genes *OSR1* (A) and *PAX6* (B). Multiple CGIs (red boxes) span the *HOXC* (C) and *PAX6* (D) gene loci. Plots of the MAP-CGI array profiles for blood, brain, muscle, and spleen identify tissue-specific CGI methylation (vertical black bars extending above $M = 1.5$). Gray bars extending downwards below $M = 1.5$ (broken blue line) represent nonmethylated CGIs. The region of *PAX6* analysed by bisulfite genomic sequencing (see Figure 5B) is indicated (asterisk in panel D). Tick marks on the *y*-axis are spaced at intervals of 1 *M* value unit. Coding sequences are diagrammed as yellow bars.

(E) Individual-specific CGI methylation internal to the *HOXC* cluster in muscle DNA.

(F) Cell type–specific methylation is seen at the *SEC31B* promoter CGI in monocytes and granulocytes derived from whole human blood. Bisulfite genomic sequencing results (A–B and E–F) are diagrammed as in Figure 3G.

doi:10.1371/journal.pbio.0060022.g005

sequence elements (TGTG/CACA) and base-stacking energy (see Figure S2). The biological relevance of these minimal differences is uncertain.

Weber and coworkers [14] identified ~2,000 promoters out of 16,000 that were more susceptible to methylation than CGIs themselves. These so-called "weak CpG islands" had an average CpG[o/e] ratio intermediate between CGIs and bulk genomic DNA. We have determined that 75% of weak CpG islands reported by Weber et al. are absent from the CGI library. Weak CGIs may be depleted because they are heavily methylated and therefore not enriched by CAP. Indeed, 22 methylated weak CpG islands [14] were not detected in our library. Alternatively, their relatively low CpG density and somewhat elevated frequency of MseI sites may result in too few CpGs per fragment for efficient retention by the CXXC matrix.

Composite Methylation of CGIs

Those CGIs that were methylated often showed a mixture of heavily methylated and nonmethylated strands by bisulfite analysis. There are several possible explanations for composite methylation patterns. Firstly, at the highest level, it is possible that different individuals contributing to the DNA pool are polymorphic with respect to this epigenetic mark. We analysed specific CGIs in muscle DNA from three individuals and found evidence of individual variation of this kind. A large-scale survey would be required to determine the extent of inter-individual variability. A second possibility is that cells within the analysed tissue are heterogeneous with respect to CGI methylation. Each of the analysed tissues consists of multiple differentiated cell types that should be analysed separately to address this possibility. Analysis of three compositely methylated CGIs in blood showed one that was highly methylated in monocytes, but weakly methylated in granulocytes, indicating that cell type–specific CGI methylation underlay heterogeneous DNA methylation. A third possible explanation for composite methylation is monoallelic methylation. A previous study of 149 CGIs on Chr 21q detected three that were monoallelically methylated, indicating that this explanation also accounts for some cases of composite CGI methylation [12].

Methods

Plasmid cloning and recombinant protein purification. Cloning of the His-CXXC construct from murine *Mbd1a* was described previously [19]. The MBD construct was subcloned from pET30bhMeCP2 [51]. A fragment of human MeCP2 corresponding to amino acids 76–167 was PCR-amplified and ligated into the NdeI and EcoR1 sites of pet30b (Novagen) to generate a C terminally His-tagged pet30b-MeCP2_76–167. Primers: pet30bMeCP2_76–167NdeI CGG TTC ATA ACC ATA TGG CTT CTG CCT CCC CCA AAC AGC GG and pet30bMeCP2_76–167EcoR1 CGG AAG TCA AAG AAT TCT CAT CAG TGG TGG TGG TGG TGC CGG GA. Recombinant peptides were purified from 10 l of induced BL21(DE3)pLysS (Stratagene) culture on Nickel Charged Fast Flow Chelating Sepharose (GE Healthcare). The CXXC construct was further purified by cation exchange using Sp-Sepharose (GE Healthcare) cation exchange as previously described [51]. Recombinant protein was bound to Nickel sepharose prior to longer term storage.

Electrophoretic mobility shift assay (EMSA). CXXC-EMSA was carried out essentially as described in [19]. Briefly, binding reactions including 0, 250, 500, 1,000, or 2,000 ng of purified recombinant His-CXXC were preincubated in 1× binding buffer (5 × binding buffer: 30 mM Tris-HCl [pH8], 750 mM NaCl, 5 mM DTT, 30 mM $MgCl_2$, 15% Glycerol, 50 ng/μl BSA, and 0.05 μg/μl of poly(dAdT) (Amersham). End-labeled CG11[52] probe (1 ng) was added to each reaction and incubated for a further 25 min. Complexes between probe DNA and the CXXC domain were resolved on a 1.3% agarose Tris-borate-EDTA gel and imaged by Phosphor Imager (Molecular Devices).

Human DNA samples. Whole blood was collected from voluntary donors and used in anonymized pools. Donors were aware of, and consented to, its use for preparation of DNA. Monocyte and granulocyte cells were prepared from whole human blood using Ficoll gradient centrifugation. Whole blood (3 ml) was layered onto an equivalent volume of Histopaque-1077 Ficoll (Sigma-Aldrich) and sedimented according to the manufacturer's instructions. Mononucleocytes were recovered from the plasma-ficoll interphase and granulocytes from the cell pellet. Whole human blood, monocyte, and granulocyte DNA was extracted using the Genomic-tip 500/G (Qiagen 10262) genomic DNA purification kit. Sperm DNA was prepared as described [53]. Human skeletal muscle, spleen, and brain genomic DNAs were purchased from Ambion.

CXXC affinity purification. 50–60 mg of recombinant CXXC was dialysed into W1 buffer (50 mM sodium phosphate buffer [pH8], 300 mM NaCl, 10% glycerol, 15 mM ß-mercaptoethanol, 0.5 mM PMSF), bound to nickel-charged sepharose, and then washed with 10 column volumes (CVs) of W1, 10 CVs of W2 (W1 + 10 mM Imadazole), and 10 CVs of W1. Beads were packed onto a 1-ml Tricorn chromatography column (GE Healthcare). MseI digested male DNA (100 μg) pooled from three individuals was bound to the CXXC column in 90% CA buffer (20 mM Hepes [pH7.9], 0.1% Triton X-100, 10% glycerol, 0.5 mM PMSF, 10 mM 2-Mercaptoethanol) and 10% CB buffer (CA + 1 M NaCl). Equilibrated DNA was then eluted over an increasing NaCl gradient of 10%–100% CB buffer (Figure 1B). Fractions (3 ml) were collected and 200 μl of each was precipitated and resuspended in 40 μl 1 × TE buffer. Aliquots were PCR- interrogated using Redhot taq DNA polymerase (Abgene) for *XIST* (for CACGTGACAAAAGCCATG, rev GGTTAGCATGGTGGTGGAC), *NYESO* (for CCCAGCGTCTGGTAACCATC, revCCACGGGACAGGTACCTC), *MAO* (for CGGGTATCAGATTGAAACAT, rev CTCTAAGCATGGCTACACTACA), *P48* (for cagaaggtcatcatctgcca, rev tgagttgtttttcatcagtcca) under the following conditions: 2 min at 94 °C; followed by 30 cycles of 94 °C for 50 s, T^{ann} °C for 50 s, 72 °C for 1 min; and a final extension of 72 °C for 7 min. PCR products were resolved on a 1.5% TAE-agarose gel (Figure 1B). Fractions retaining nonmethylated CpG-rich MseI fragments (Figure 1B) were pooled, diluted with CA buffer, and re-chromatographed. The relevant fractions were precipitated and ligated into the NdeI site of pGEM5zf- (Promega).

CGI library sequencing. The clone set was arrayed into 384-well plates in glycerol for long-term storage. Copies were taken and DNA prepared for sequencing using a modified alkaline lysis method. Cells were lysed in glucose, Tris, EDTA (pH 8) buffer plus NaOH and SDS and spun through Millipore Montage filter plates directly into propan-2-ol to precipitate, followed by elution in water. In all, 172,800 clones were sequenced forward and reverse using T7 and SP6 primers and BigDye V3.1 chemistry, under the following conditions: 30 s at 96 °C, followed by 44 cycles of 92 °C 8 s, 55 °C 8 s, 60 °C 2 min. Samples were separated using 3730 XL sequencers (Applied Biosystems). Extraction was performed using sequence analysis v3.1, and base-called using Phred [54]. DNA sequences were identified using NCBI36 and mapped using the ENSMBL Genome Browser (http://www.ensembl.org/index.html). CGIs that mapped within 1.5 kb of annotated genes were considered to be gene-associated in order to

take into account mis-annotation of transcription start sites within poorly defined 5′ UTRs.

Microarray fabrication. Amino-linked clone insert amplicons were generated by vector-specific PCR in 50 mM KCl, 5 mM Tris (pH 8.5), and 2.5 mM $MgCl_2$ including 1 M Betaine (10 min at 95 °C; followed by 35 cycles of 95 °C for 1 min, 60 °C for 15 s, 72 °C for 7 min; and a final extension of 92 °C for 10 min; 5′ aminolink forward primer 5′ -CTC ACT ATA ggg CgA ATTg g –3′ reverse primer 5′ -CgC CAA gCT ATT TAg gTg AC-3′). PCR products were ethanol-precipitated and resuspended in 1 × microarray spotting buffer (250 mM sodium phosphate [pH 8.5], 0.01% sarkosyl, 0.1% sodium azide). Arrays were spotted onto amine-binding slides at 20–25 °C, 40%–50% relative humidity. After an overnight incubation in a humid chamber, the slides were blocked (1% ammonium hydroxide for 5 min, followed by 0.1% SDS for 5 min) and denatured (95 °C ddH_2O for 2 min), rinsed in ddH_2O, and dried by centrifugation for 5 min at 250 ×*g*.

MAP, labeling, and microarray hybridization. Human tissue DNA pooled from three individuals was digested with MseI, phosphatase-treated, and ligated to 5 μmol of phosphorylated catch-linkers (upper__GGT CCA TCC AAC CGA TCT and lower__CCA GGT AGG TTG GCT AGA AT phosphate) that had been annealed in 1× TE for 5 h. DNA was bound to an MBD chromatography column and affinity-purified essentially as described [20]. Fractions containing methylated CpG-rich MseI fragments were pooled and re-chromatographed before precipitation (Figure 3B). Purified DNA was resuspended in 1×TE and amplified in parallel with input DNA, using the GC Rich PCR system (Roche; 2 min at 95 °C; followed by 18 cycles of 95 °C for 1 min, T^{ann} °C for 1 min, 72 °C for 4 min; and a final extension of 72 °C for 7 min; universal primer__GGT CCA TCC AAC CGA TCT TA). MAP and Input DNAs (200 ng) were fluorescently labeled by random priming using the Bioprime labeling kit (Invitrogen), 1 × dNTS (10× dNTPS; 2 mM of each dATP, dGTP, dTTP, 1 mM dCTP) and 1.5 nmol of Cy3 or Cy5-dCTP (GE Healthcare). The labeled Input and MAP probes were purified (Invitrogen "Purelink"), pooled, and precipitated with 100 μg of human Cot-1 DNA (Invitrogen). Labeled DNA was resuspended in 400 μl of hybridisation buffer (2XSSC, 50% deionised formamide, 10 mM Tris-HCl [pH7.5], 10% dextran sulphate, 0.1% Tween 20), denatured at 100 °C for 10 min, snap-chilled on ice, and incubated for 1 h at 37 °C. The CGI microarrays were prehybridised with Cot-1 and herring sperm DNA (Sigma) before being hybridised for 48 h at 37 °C. Arrays were washed four times at 37 °C in 1× phosphate buffered saline/0.05% Tween 20, three times at 52 °C in 1 × saline sodium citrate, twice at RT in 1 × phosphate buffered saline/0.05% Tween 20, and finally rinsed in water, before being dried by centrifugation (500*g*).

Microarray scanning and data analysis. Arrays were scanned with a GenePix Autoloader 4200AL (Axon) and then processed using the GenePix Pro 6.0 (Axon) software package. All subsequent analysis was carried out with the LIMMA package in the *R* statistical environment. Features with poor signal-to-noise ratios were stabilised using a base value of 1,000 for background-subtracted intensities. Cy3 and Cy5 signals were transformed into *M* values ($\log_2$[red/green]) and normalised by print-tip loess. Each tissue analysis is represented by four microarrays comprising two independent replicates with respective dye swaps. Processed values were averaged through linear modeling and used to determine the relative enrichment of MAP DNA relative to Input. An *M* value of >1.5 was designated as the threshold for hypermethylation as determined by quantitative PCR (Figure 3C) and bisulfite genomic sequencing (Figures 3G and 3H, 4D–4G, and 5A and 5B). This threshold was confirmed as significant by calculation of a *t*-statistic by eBayes modeling and BH multiple testing correction. Differential methylation was deduced when features displayed an *M* value >1.5 in one or more tissues and a differential of 0.75 between tissues (upper boundary capped at $M = 2.5$). To avoid complications due to X chromosome inactivation, CGIs on sex chromosomes were not included in the analysis. In addition, spots that gave no signal on the microarray (NA values) and spots containing DNA in which CpG[o/e] values were <0.5 were excluded.

Quantitative PCR. Real-time PCR was carried out on MAP and Input material with iQ SYBR Green Supermix (Bio-Rad) on an iCycler (Bio-Rad) according to manufacturer's instructions. For primer sequences see Table S4.

Bisulfite genomic sequencing. Bisulfite treatment of genomic DNA was carried out as described by Feil et al. [55], and prepared for sequencing as outlined by Suzuki et al. [56]. Genomic DNA (5 μg) was digested by EcoRI prior to bisulfite treatment, and precipitated after the desulfonation step. Samples were resuspended in 1 ×Tris-EDTA buffer for subsequent PCR and sequencing reactions. Bisulfite specific primers were designed both manually and with the aid of the MethPrimer software [3] (sequences are available on request). PCR was carried out on the bisulfite-treated DNA using RedHot Taq DNA polymerase (Abgene) under the following conditions: 2 min at 94 °C; followed by 40 cycles of 94 °C for 50 s, T^{ann} °C for 50 s, 72 °C for 1 min; and a final extension of 72 °C for 5 min. PCR fragments were cloned using the Strataclone PCR cloning system (Stratagene) and at least ten products amplified (as above) and sequenced (BigDye Terminator v3.1 Cycle Sequencing Kit; Applied Biosystems). Methylation status and experimental quality control was carried out with the aid of BiqAnalyzer [57].

Supporting Information

Dataset S1. Characterization of Human CGIs Identified as Being Methylated in One or More Somatic Tissues

Included are sequence characteristics (G+C and CpG[o/e]), gene association (Gene ID, ENSEMBL nomenclature), and the methylation profile in the four tissues tested (categories of methylation are colour-coded as for Figure 4B). The CpG island identifier (ID) corresponds to the arrayed CGI fragment. The genomic position of each CGI (Location) corresponds to the chromosomal coordinates derived from NCBI build 36.

Found at doi:10.1371/journal.pbio.0060022.sd001 (373 KB XLS).

Figure S1. Sequence Characteristics of CGIs Missed by NCBI-Strict

NCBI-strict relies on base composition to identify CGIs, utilising threshold values for CpG[o/e] and G+C density (0.6% and 50%, respectively) as determinants. Boxplots of G+C and CpG[o/e] indicate that CGIs retained by the CXXC affinity matrix but missed by NCBI-strict have significantly reduced G+C base composition (p-value < $2.2e^{-16}$) and CpG[o/e] (p-value < $2.2e^{-16}$). A nonparametric distribution was determined using a Shapiro-Wilk test of normality and subsequent significance was determined using a two-sample Kolmogorov-Smirnov test (NCBI-missed $n = 4,082$ and all CGIs $n = 13,305$).

Found at doi:10.1371/journal.pbio.0060022.sg001 (48 KB DOC).

Figure S2. Sequence Properties of Methylated CGIs

Bock and colleagues determined a number of DNA sequence features that are correlated with DNA methylation at CGIs [50]. Here we compare the sequence attributes of the methylated and total CGI sets with respect to DNA structure (stacking energy and base twist) and specific repeats (TGTG/CACA). Methylated CGIs show small but significant increase in stacking energy relative to all CGIs (p-value < 0.001). In contrast we found no significant difference in the base twist of methylated CGIs.

TGTG/CACA specific repeats were found to be significantly enriched in methylated CGIs (p-value < 0.001; see text for discussion). In contrast, all repetitive elements (as outlined in Repbase [58]) were found to be marginally depleted in methylated CGIs (p-value < 0.01, Wilcoxon rank sum test, $n = 4,082$ and 10,236). Stacking energy and base twist were calculated using the EMBOSS b-twisted program with default settings[59]. All distributions were tested for parametric distribution by the Shapiro-Wilk test of normality. Nonparametric significance values were determined using the Wilcoxon rank sum test ($n = 4,082$ and 10,236).

Found at doi:10.1371/journal.pbio.0060022.sg002 (66 KB DOC).

Table S1. Gene Association: CpG Islands Missed by NCBI-Strict

All CGIs ($n = 4,082$) retained by the CXXC affinity matrix but not predicted by NCBI-strict were mapped relative to protein-coding genes. Gene overlap indicates the spatial association of CGIs relative to protein-coding genes.

Found at doi:10.1371/journal.pbio.0060022.st001 (27 KB DOC).

Table S2. Methylated CGIs on Chromosome 16 and Chromosome X in Human Whole Blood DNA

CGI arrays hybridised with MBD probe from male and female human blood DNA identify extensive methylation of X-linked CGIs. In contrast, the percentage of X-linked CGIs methylated on the single male X chromosome was comparable to the levels found on the human autosomes, as illustrated for Chr16.

Found at doi:10.1371/journal.pbio.0060022.st002 (30 KB DOC).

Table S3. Developmental Gene Categories Are Associated with Differentially Methylated CpG Islands

Ontology terms for gene-associated CGIs were compared with those for differentially methylated CGI-genes. Genes involved in developmental processes such a neurogenesis and segmentation are

significantly enriched and include transcriptional regulators such as homeobox genes. Significantly enriched biological processes and molecular functions were determined using the Web-based Panther classification system (http://www.pantherdb.org/)[60].

Found at doi:10.1371/journal.pbio.0060022.st003 (41 KB DOC).

Table S4. Quantitative PCR Primers for Microarray Validation

CGI ID: CpG island identifiers which correspond to the CGI library. *M* value: Log_2[MBD/Input] for human blood microarray experiment.

Found at doi:10.1371/journal.pbio.0060022.st004 (36 KB DOC).

Acknowledgments

We are grateful to Anne Evans and Aileen Greig for excellent technical support; to Jose de las Heras, Robert Andrews, Ian Still, Michael Quail, and Carol Scott for help and advice; and to Miho Suzuki, Heather Owen, and Pete Skene for comments on the manuscript. We thank the Wellcome Trust Sanger Institute Microarray Facility and Mapping Core Group.

Author contributions. AB, RI, and HJ conceived and designed the experiments. RI and DD performed the experiments. RI, AK, JS, DJ, and TC analyzed the data. HJ contributed reagents/materials/analysis tools. PE, SH, and CL designed and printed the CGI microarrays. CC, RP, and JR performed DNA sequencing of the CGI library. RI and AB wrote the paper.

Funding. This work was supported by the Wellcome Trust and by a Medical Research Council (UK) studentship to RI.

Competing interests. The authors have declared that no competing interests exist.

References

1. Li E, Bestor TH, Jaenisch R (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell 69: 915–926.
2. Okano M, Bell DW, Haber DA, Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99: 247–257.
3. Bird A (2002) DNA methylation patterns and epigenetic memory. Genes Dev 16: 6–21.
4. Bird A, Taggart M, Frommer M, Miller OJ, Macleod D (1985) A fraction of the mouse genome that is derived from islands of non-methylated, CpG-rich DNA. Cell 40: 91–99.
5. Stein R, Razin A, Cedar H (1982) In vitro methylation of the hamster adenine phosphorybosy transferase gene inhibits its expression in mouse L cells. Proc Natl Acad Sci U S A 79: 4418–3422.
6. Hansen RS, Gartler SM (1990) 5-Azacytidine-induced reactivation of the human X chromosome-linked PGK1 gene is associated with a large region of cytosine demethylation in the 5′ CpG island. Proc Natl Acad Sci U S A 87: 4174–4178.
7. Heard E, Clerc P, Avner P (1997) X-chromosome inactivation in mammals. Annu Rev Genet 31: 571–610.
8. Sado T, Fenner MH, Tan SS, Tam P, Shioda T, et al. (2000) X inactivation in the mouse embryo deficient for Dnmt1: distinct effect of hypomethylation on imprinted and random X inactivation. Dev Biol 225: 294–303.
9. Reik W (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. Nature 447: 425–432.
10. De Smet C, Lurquin C, Lethe B, Martelange V, Boon T (1999) DNA methylation is the primary silencing mechanism for a set of germ line and tumor-specific genes with a CpG-rich promoter. Mol Cell Biol 19: 7327–7335.
11. Strichman-Almashanu LZ, Lee RS, Onyango PO, Perlman E, Flam F, et al. (2002) A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. Genome Res 12: 543–554.
12. Yamada Y, Watanabe H, Miura F, Soejima H, Uchiyama M, et al. (2004) A comprehensive analysis of allelic methylation status of CpG islands on human Chromosome 21q. Genome Res 14: 247–266.
13. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, et al. (2006) DNA methylation profiling of human Chromosomes 6, 20, and 22. Nat Genet 38: 1378–1385.
14. Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, et al. (2007) Distribution, silencing potential, and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 39: 457–466.
15. Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A 103: 1412–1417.
16. Gardiner-Gardner M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196: 261–282.
17. Takai D, Jones PA (2003) The CpG island searcher: a new WWW resource. In Silico. Biol 3: 235–240.
18. Voo KS, Carlone DL, Jacobsen BM, Flodin A, Skalnik DG (2000) Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. Mol Cell Biol 20: 2108–2121.
19. Jorgensen HF, Ben-Porath I, Bird AP (2004) Mbd1 is recruited to both methylated and nonmethylated CpGs via distinct DNA binding domains. Mol Cell Biol 24: 3387–3395.
20. Cross SH, Charlton JA, Nan X, Bird AP (1994) Purification of CpG islands using a methylated DNA binding column. Nat Genet 6: 236–244.
21. Kochanek S, Renz D, Doerfler W (1993) DNA methylation in the Alu sequences of diploid and haploid primary human cells. EMBO J 12: 1141–1151.
22. Brock GJ, Charlton J, Bird A (1999) Densely methylated sequences that are preferentially localized at telomere-proximal regions of human chromosomes. Gene 240: 269–277.
23. Takai D, Jones PA (2002) Comprehensive analysis of CpG islands in human Chromosomes 21 and 22. Proc Natl Acad Sci U S A 99: 3740–3745.
24. Antequera F, Bird A (1993) Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A 90: 11995–11999.
25. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, et al. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet 37: 853–862.
26. Carrel L, Clemson CM, Dunn JM, Miller AP, Hunt PA, et al. (1996) X inactivation analysis and DNA methylation studies of the ubiquitin activating enzyme E1 and PCTAIRE-1 genes in human and mouse. Hum Mol Genet 5: 391–401.
27. Carrel L, Willard HF (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434: 400–404.
28. Futscher BW, Oshiro MM, Wozniak RJ, Holtan N, Hanigan CL, et al. (2002) Role for DNA methylation in the control of cell type–specific maspin expression. Nat Genet 31: 175–179.
29. Shiota K, Kogo Y, Ohgane J, Imamura T, Urano A, et al. (2002) Epigenetic marks by DNA methylation specific to stem, germ, and somatic cells in mice. Genes Cells 7: 961–969.
30. Song F, Smith JF, Kimura MT, Morrow AD, Matsuyama T, et al. (2005) Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. Proc Natl Acad Sci U S A 102: 3336–3341.
31. Nguyen C, Liang G, Nguyen TT, Tsao-Wei D, Groshen S, et al. (2001) Susceptibility of nonpromoter CpG islands to de novo methylation in normal and neoplastic cells. J Natl Cancer Inst 93: 1465–1472.
32. Kodama R, Eguchi G (1994) Gene regulation and differentiation in vertebrate ocular tissues. Curr Opin Genet Dev 4: 703–708.
33. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, et al. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell 126: 1189–1201.
34. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet 39: 61–69.
35. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823–837.
36. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130: 77–88.
37. Scherer SE, Muzny DM, Buhay CJ, Chen R, Cree A, et al. (2006) The finished DNA sequence of human Chromosome 12. Nature 440: 346–351.
38. Ooi SK, Qiu C, Bernstein E, Li K, Jia D, et al. (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature 448: 714–717.
39. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196: 261–282.
40. Macleod D, Ali RR, Bird AP (1998) An alternative promoter in the mouse major histocompatibility cmplex class II I-Aβ gene: implications for the origin of CpG islands. Mol Cell Biol 18: 4433–4443.
41. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, et al. (2004) Global identification of human transcribed sequences with genome tiling arrays. Science 306: 2242–2246.
42. Panning B, Jaenisch R (1996) DNA hypomethylation can activate Xist expression and silence X-linked genes. Genes Dev 10: 1991–2002.
43. Sleutels F, Zwart R, Barlow DP (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature 415: 810–813.
44. Wutz A, Smrzka OW, Schweifer N, Schellander K, Wagner EF, et al. (1997) Imprinted expression of the Igf2r gene depends on an intronic CpG island. Nature 389: 745–749.
45. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129: 1311–1323.
46. Rauch T, Wang Z, Zhang X, Zhong X, Wu X, et al. (2007) Homeobox gene

methylation in lung cancer studied by genome-wide analysis with a microarray-based methylated CpG island recovery assay. Proc Natl Acad Sci U S A 104: 5527–5532.
47. Ohm JE, McGarvey KM, Yu X, Cheng L, Schuebel KE, et al. (2007) A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat Genet 39: 237–242.
48. Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M, et al. (2007) Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat Genet 39: 232–236.
49. Widschwendter M, Fiegl H, Egle D, Mueller-Holzner E, Spizzo G, et al. (2007) Epigenetic stem cell signature in cancer. Nat Genet 39: 157–158.
50. Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, et al. (2006) CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. PLoS Genet 2: e26. doi:10.1371/journal.pgen.0020026
51. Klose RJ, Bird AP (2004) MeCP2 behaves as an elongated monomer that does not stably associate with the Sin3a chromatin remodeling complex. J Biol Chem 279: 46490–46496.
52. Meehan RR, Lewis JD, McKay S, Kleiner EL, Bird AP (1989) Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. Cell 58: 499–507.
53. Hossain AM, Rizk B, Behzadian A, Thorneycroft IH (1997) Modified guanidinium thiocyanate method for human sperm DNA isolation. Mol Hum Reprod 3: 953–956.
54. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8: 175–185.
55. Feil R, Charlton J, Bird AP, Walter J, Reik W (1994) Methylation analysis on individual chromosomes: improved protocol for bisulphite genomic sequencing. Nucleic Acids Res 22: 695–696.
56. Suzuki MM, Kerr AR, De Sousa D, Bird A (2007) CpG methylation is targeted to transcription units in an invertebrate genome. Genome Res 17: 625–631.
57. Bock C, Reither S, Mikeska T, Paulsen M, Walter J, et al. (2005) BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. Bioinformatics 21: 4067–4068.
58. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462–467.
59. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276–277.
60. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13: 2129–2141.